Understanding question-reading deviations: Implications for monitoring interviewers, questionnaire design, and data quality

Jennifer Kelley

A thesis submitted for the degree of

Doctor of Philosophy in Survey Methodology

Institute for Social and Economic Research

University of Essex

December 2020

Declarations

No part of this has been submitted for another degree.

I am the sole author of Chapter 1 and Chapter 2. Chapter 3 is co-authored with Tarek Al Baghal, University of Essex. I did the behavior coding, data management and wrote the first draft of the paper. Tarek and I worked together on the analysis and editing the paper.

A version of Chapter 1 has been published as:

> Kelley, J. (2020). Accuracy and Utility of Using Paradata to Detect Question-Reading Deviations. *Interviewer Effects from a Total Survey Error Perspective*, 267.

I dedicate this to Brad, Danielle, and Zach.

Acknowledgments

I am deeply grateful to the people who encouraged and supported me through my education journey. I wish to thank my supervisor Tarek Al Baghal for his unwavering support these past four years. His invaluable feedback and guidance have allowed me to expand my research skills, which in turn has provided me a solid foundation to build upon for my future research endeavors. I also wish to thank my second supervisor, Peter Lynn, for his steadfast support and feedback. I am thankful for the opportunity to learn from such accomplished survey methodologists.

I am very grateful for the opportunity to study at the prestigious Institute for Social and Economic Research with the financial support of ESRC. As an American studying in the United Kingdom, the people of ISER graciously welcomed me. I would especially like to thank Tarek and Cara Booker, two fellow Americans, who helped me navigate the differences between the U.S. and the U.K. academic worlds. During my time at ISER, I enjoyed many discussions on survey methodology with my fellow PhD students, Alex Wenz, Linh Nguyen, Valerija Kolbas, and Brendan Read. I treasured our student-initiated weekly meetings; it was a safe and supportive atmosphere to discuss our research ideas and receive feedback on our work.

I am also very grateful to several people in my professional life who supported me through my PhD studies. Thank you, Robby Stewart, for being my first academic mentor, teaching me to love research, and how to prepare for conference presentations. Thank you, Beth-Ellen Pennell, for taking a chance on a non-traditional student as your graduate student research assistant, and your invaluable mentorship throughout the years. I would also like to thank several people at the University of Michigan, Zeina Mneimneh, Nicole Kirgis, Gina Cheung, and Nancy Bylica, who

supported my academic break from the Survey Research Center as I pursued my academic goals. I cherish their support and friendship.

I want to thank my family and friends who have supported me every step of the way. To my late father-in-law, Dan, who pushed me to take the first step as a non-traditional student - you will never know how that push changed my life. To my sister, Meribeth, who keeps me sane and will always be the Lucy to my Ethel. To my mom, who taught me that women could do anything. To my children, Danielle and Zach, who never complain about having an eternal student as a mom and always cheer me on. To my granddaughters, Nora and Jena, your unconditional love fills me up and inspires me to be the best role model I can be. Finally, to my husband, Brad, who is my biggest cheerleader. Thank you for clearing the path at home so that I could pursue my dreams. I cannot express how grateful I am for your love and support over the past 30 years.

Summary

Chapter 1 examines the accuracy and utility of using paradata to detect interviewer question-reading deviations. Using timestamps and behavior coded data from interviewer recordings, I explore different methods (i.e., different rates and ranges of reading pace, standard deviations, and model-based methods) for constructing question administration timing thresholds (QATT) and compare them to the behavior coded data to determine the accuracy and utility of each method to detect minor and major deviations. Results show that using a reading rate of 4 words per second (WPS) to create upper and lower QATTs has the highest overall accuracy (87.1%) and the most utility for correctly identifying interviews with and without major deviations.

Chapter 2 examines the impact question characteristics have on question-reading deviations in face-to-face interviews. To evaluate this, questions from the Innovation Panel (IP) Wave 3 were coded on the following dimensions: structure, content, and the presence of interviewer aids, resulting in 19 question characteristics. Results show that of the 19 question characteristics examined, 16 are significantly associated with major question-reading deviations. The question characteristics that have the highest odds of major deviations are questions that have definitions or examples (6.404), questions that have response options read in the question text (4.133), and demographic questions (2.421).

Chapter 3 examines the impact of question-reading deviations on data quality. Several measures are used to assess data quality, including item non-response and differences in response distributions for questions that are read verbatim (or have minor deviations) and questions that have major deviations. The results show that major question-reading deviations are only significantly associated with question timing; changed wording has a significant negative

association with question timing. The other data quality indicators (i.e., Don't Know and distribution of means) showed no significant effect from major question-wording deviations.

Table of Contents

List of Figures

List of Tables

Introduction

As survey research has evolved and advanced over the years, face-to-face surveys have remained the primary data collection mode for many large national and international household surveys. Some household surveys have adapted to mixed mode, but face-to-face remains part of the equation as either the primary mode or as a follow-up mode for respondents who do not complete the interview in less expensive mode offered (e.g., telephone, web, mail). The reason for face-to-face staying power is that it has long been the gold standard to which other modes of data collection are compared, despite concerns about interviewer effects and rising costs.

Interviewers' roles have also evolved (e.g., using new technologies, collecting bio measures, administering mental and physical tests), but the technique for how interviewers administer questions has remained the same; standardized interviewing. Standardized interviewing techniques are widely used as they have been shown to reduce interviewer effects or measurement error (Groves, Fowler, Couper, Lepkowski, Singer & Tourangeau, 2011; Krosnick, Malhotra, and Mittal, 2014). The cornerstone of standardized interviewing is reading questions as written, verbatim. However, it is well-documented that interviewers do not always read questions verbatim (Ackermann-Piek and Massing, 2014; Cannell, Lawson, & Huasser, 1975; Mathiowetz & Cannell, 1980), hence organizations are encouraged to monitor how interviewers read questions. The methods for monitoring question-reading has also largely remained the same – recording and listening to interviews, or observing (face-to-face interviews), or listening (telephone) to real-time interviews. Both of these methods are resource and time-intensive, so if organizations monitor question-reading behavior, they only monitor the first few interviews and randomly select a small percentage of subsequent interviews (Thissen & Myers, 2016; Viterna &

Maynard, 2002). Advances in survey software allow organizations to monitor interviewers' behavior using paradata, which *may* significantly reduce the resources needed to monitor question-reading behavior, but little is known about the methods' accuracy and utility.

Detecting interviewer deviations from the questionnaire script is important for making sure interviewers follow protocol, but knowing more about what is driving interviewers to deviate would enable researchers, questionnaire designers, and interviewer trainers to take proactive steps to stop or greatly reduce deviations. Is it the characteristics of the question, the respondent or the interviewer driving the behavior, or some combination of the three characteristics? The studies that have attempted to identify the source of what is driving the behavior have been limited in scope, in terms of question types or respondent and interviewer characteristics (Bradburn, Sudman, Blair, Locander, Miles, Singer & Stocking, 1979; Cannell & Robison, 1971; Mathiowetz & Cannell, 1980), or used telephone data (Presser and Zhao, 1992), which have been shown to have much fewer deviations (Ackermann-Piek and Massing, 2014; Cannell, Lawson, & Huasser, 1975). There is no known face-to-face study that evaluates an extensive list of question characteristics and question reading deviations.

Evaluating the most efficient way to detect question reading deviations and the mechanisms driving the behavior are essential steps for reducing deviations. However, what may be even more important to know is how deviations impact data quality. Here the literature is even more sparse, and the findings are mixed; some find a negative association with data quality (Schumann and Presser, 1996), others find a positive association (Dykema, Lepkowski, and Blixt, 1997; Haan, Ongena, and Huiskes, 2013) and still others find a mix of positive and negative

associations (Belli, Lee, Stafford, and Chou, 2004). These results suggest that deviations may have a differential effect depending on the type of question, but more research is needed.

This thesis uses a unique data set consisting of paradata, survey data, and behavior coded data derived from interview recordings from Wave 3 of the Understanding Society Innovation Panel (IP). This unique data set provides the opportunity to evaluate the three areas discussed: 1) using paradata to monitor interviewers' question-reading behavior, 2) the role of question characteristics in interviewer deviations, and 3) deviations and data quality. The three chapters are outlined below.

Chapter 1 uses paradata and behavior coded data from interviewer recordings to explore different methods (i.e., different rates and ranges of reading pace, standard deviations, and model-based methods) for constructing question administration timing thresholds (QATT) to detect minor and major deviations. The question timing durations (derived from paradata) are compared to the QATTs to identify questions that violate the questions' QATTs, and violations are flagged as possible question-reading deviations. The QATT violations are then compared to the behavior coded data (i.e., how the interviewers *actually* administered the question) to evaluate the accuracy of the different QATT detection methods. The data is then aggregated to the interview level to assess each of the QATT detection methods' utility.

Chapter 2 focuses on the impact question characteristics have on question-reading deviations in face-to-face interviews. Specifically, are there certain types of questions that have a higher probability of interviewers making question-reading deviations? To evaluate this, questions from the IP Wave 3 were coded on the following dimensions: structure, content, and the presence of interviewer aids, resulting in 19 question characteristics. The relationship between the question

characteristics and interviewers' deviations are first assessed using bivariate analysis. A multilevel logistic regression model with question, respondent, interviewer, and interview context level variables is used to explore the relationship in more depth.

Chapter 3 uses behavior coded data, timing data, and survey data to evaluate question-reading deviations and data quality. Several measures are used to assess data quality, including item non-response and differences in distributions for questions that are read verbatim (or have minor deviations) and questions that have major deviations. In addition, this study exploits several IP Wave 3 experiments on question formation (e.g., branching and presence of showcards) to evaluate whether or not the measurement error (i.e., differential response distributions) found for different question formations can be partially attributed to interviewer question-reading deviations.

Accuracy and Utility of Using Paradata to Detect Question-Reading Deviations

Abstract

Deviations from reading survey questions exactly as worded may change the validity of the questions, thus increasing measurement error. Hence, organizations train their interviewers to read questions verbatim. To ensure interviewers are reading questions verbatim, organizations rely on interview recordings. However, this takes a significant amount of resources. Therefore, some organizations are using paradata generated by the survey software, specifically timestamps, to try to detect when interviewers' deviate from reading the question verbatim. However, there is no established method on how to use timestamps to detect question-reading deviations, and little is known about the level of accuracy for the different methods currently used.

This study evaluates the different methods for detecting question-reading deviations using interview recordings and paradata from Wave 3 of the Understanding Society Innovation Panel. Using interview recordings allows a direct comparison of the different detection methods to how the interviewers *actually* administered the question and thus measures each detection method's accuracy and utility. Deviations will also be coded for the extent (i.e., minor or major) and type of deviation. This analysis will give better insight into the scope and types of deviations interviewers engage in and practical guidance on how to best detect question-reading deviations.

1.1 Introduction

Data are everywhere. From wearables tracking each step a person takes, to thermostats tracking household heating preferences, to social media capturing internet browsing history, there is a plethora of data. Survey research is no exception. Advances in survey software, including managing samples and conducting interviews, can now capture the survey's *process* data at every stage of the survey lifecycle, creating substantial amounts of data. This micro-level process data are known in the survey world as paradata (Kreuter, 2013). Paradata are appealing to survey organizations because the data can be captured with relative ease and at little or no cost. Paradata has the promise of reducing study costs while improving field-operation efficiency and data

quality. Hence, survey organizations use paradata throughout the survey lifecycle, from study design to field operations to post-survey adjustments.

Focusing on the field operations phase, organizations are using paradata in several ways, including monitoring interviewers' behavior. They use paradata, like keystrokes and timestamps, to monitor interviewers' behavior to detect issues with interviewers' performance or issues with the questionnaire or instrument. For example, if interviewers frequently use a 'help' key on a given question, this action could indicate a problem with respondent comprehension or a technical issue with the instrument for that question. Analyzing keystroke paradata allows researchers to not only detect issues with the questionnaire or survey protocols but it also allows them to evaluate the magnitude of the issue. Researchers can then make informed decisions on how to intervene best or address the issues based on empirical evidence, not anecdotal evidence.

The potential power of paradata is propelling organizations to look for new ways to leverage paradata to improve survey operations and data quality. While timestamps, or more accurately timing durations, have been used relatively early on in the paradata revolution to calculate interview lengths (e.g., aggregating timing durations to the interview level) and to detect respondent comprehension issues with individual questions, a new trend is starting to emerge that uses timing durations to monitor interviewers' behavior during the interview and evaluate measurement error (i.e., data quality).

Organizations that use timing durations to monitor interviewers use the timing durations as a proxy for how interviewers read questions. To reduce measurement error, organizations train

their interviewers to read the question precisely as worded, so each respondent receives the same stimuli. Deviations from reading the question exactly as worded may change the question's validity, thus increasing measurement error (Groves et al., 2011; Krosnick, Malhotra, and Mittal, 2014). To monitor interviewers' question reading behavior, organizations estimate the *expected* question administration time to establish a minimum and maximum question administration time thresholds (QATT). They then compare the question timestamp to the QATTs to identify questions that violate the question's QATTs. Violations of minimum QATTS may indicate interviewers omitted words from the question text.

Conversely, violations of maximum QATTs may indicate interviewers added words to the question text.[1] The QATT violated questions are then flagged for further investigation. Investigations may include such things as listening to the recording for a said question or aggregating the data (i.e., the flagged questions) up to the interviewer level to identify interviewers who repeatedly engage in question-reading deviations. Organizations can then make decisions about training needs or disciplinary actions based on empirical data. However, there is no established way to calculate QATTs. Some organizations calculate QATTs by dividing the question words by an (x) reading pace (Sun & Meng, 2014) or a priori cutoff, such as one second (Mneimneh, Pennell, Lin, & Kelley, 2014).

Further, there is little known about the accuracy of the methods currently used to detect question-reading deviations or if a more accurate method is needed. Which QATT method is more accurate for detecting questions that were not read verbatim? Should one construct QATTs using

[1]Interviewers may also substitute words in the question text and is discussed in the Background section.

words per second (WPS) or use standard deviations of the mean reading-time? What WPS rate or standard deviation should be used? Is one detection method better for detecting certain types of deviations (e.g., skipping words or questions, adding words to the question)?

This study will take advantage of a unique data set from Wave 3 of the Understanding Society Innovation Panel, including question timing paradata and behavior coded data from interview recordings. Using interview recordings allows a direct comparison of the different detection methods to how the interviewers *actually* administered the question and measures each detection method's accuracy. In addition, the interview recordings will be coded for the extent (i.e., minor or major) and type of deviation. This analysis will give better insight into the scope and types of deviations interviewers engage in and practical guidance on how to best detect deviations.

1.2 Background

Interviewers' behavior and measurement error

Interviewer characteristics (e.g., race, gender) and how the interviewer behaves during the interview process contribute to measurement error (Axinn, 1991; Groves, Fowler, Couper, Lepkowski, Singer & Tourangeau, 2011). In an attempt to lower the interviewer effect, organizations engage in several design strategies, including, but not limited to, recruiting interviewers to match interviewers and respondents on specific characteristics (e.g., race or gender) as the characteristics relate to the topics of the interview, training interviewers to strictly follow study protocols, act in a professional and neutral manner, balancing interviewer workloads (i.e., not assigning an interviewer a disproportionate amount of interviews), and supervising and monitoring interviews for quality issues.

Survey interviews do not follow the norms of everyday conversation (Houtkoop-Steenstra, 2000). Therefore interviewers are trained in interviewing techniques. The two interviewing methods most widely used in survey research are standardized interviewing and, to a lesser extent, conversational interviewing. Interviewers trained in standardized interviewing are instructed to maintain a professional and neutral manner throughout the interview process, read the questions verbatim and in order as they appear in the instrument, and address any issues of comprehension with a scripted set of probes (e.g., which is closer [to how many times you visited the doctor, '1' or '2']) and responses to respondent questions (e.g., whatever it means to you), so that each respondent receives the same stimuli.

For most surveys, considerable efforts have been put into developing a valid questionnaire. Questions can be subjected to any or all of the following questionnaire development methods: expert reviews, focus groups, cognitive testing, and pilot testing. Substantive experts and survey methodologists can spend months drafting, testing, and revising questions, often with several iterations of this process, to produce a valid and sound questionnaire. One reason for this is researchers have learned changes in question-wording can change the meaning of the question (Groves et al., 2011; Krosnick, Malhotra, & Mittal, 2014; Schuman & Presser, 1996). Hence, when interviewers deviate from reading the question exactly as worded, depending on the gravity of the deviation, they may be changing the meaning and validity of the question, thus increasing measurement error.

However, it is unrealistic to think every interviewer reads every question exactly as worded every single time. Published estimates of how often interviewers deviate from reading questions

exactly as worded are difficult to find, as investigations are expensive and often done in-house, and the results are proprietary. Of the sparsely published literature, several studies conducted in telephone labs estimate the rate of question-reading deviation taken by interviewers to be as low as 4.6% (Mathiowetz & Cannell, 1980) to as high as 36% (Cannell, Lawson, & Huasser, 1975). Cannell et al. (1975) go on to state, "20% [of questions] were altered sufficiently to destroy comparability". In face-to-face interviews, where supervisors are removed from the workspace (i.e., respondents' homes), one study estimates the rate of interviewers not reading the questions exactly as worded is as high as 84% (Ackermann-Piek and Massing, 2014).

Deviations can come in many forms. Interviewers are human and thus make simple human errors in reading the question (e.g., substituting 'the' for 'a'). Other interviewers may intentionally change the wording because they think they are 'helping' respondents comprehend the question (Schober & Conrad, 2002). Also, interviewers may 'tailor' the question to the respondent to signal they have listened to the respondent's previous answers (Ongena & Dijkstra, 2006b). In more extreme cases, interviewers may shorten questions by omitting words or skip questions entirely to shorten the overall interview length.

While there are no known studies of *why* interviewers do not read questions exactly as worded, one could argue motivations may be altruistic or selfish. Altruistic motivations may stem from the interviewer picking up cues from the respondent that they are tired (e.g., respondent asks "how much longer?") or frustrated (e.g., respondent asks "didn't you ask me this question already?"). Consequently, the interviewer tries to 'speed up' the interview and avoid an interview break-off (i.e., stopping the interview before it is finished) by omitting question text or

skipping questions altogether, thinking they are sacrificing item nonresponse, or measurement error, for a completed interview. Motivations that may be less altruistic can range from interviewers simply not wanting to put in the effort required by the study's protocols to wanting to speed up the interviewer process for their own reasons (e.g., they are becoming impatient with a respondent who digresses frequently, they do not like the respondent). The pay structure may also contribute to why interviewers engage in shortcuts. Interviewers that are paid per interview, are more likely to be incentivized to have short, quick interviews so that they can complete more interviews than those paid by a per-hour pay structure. Regardless of the type of deviation or the motivation for doing so, interviewers who do not read questions precisely as worded jeopardize the questions' validity.

Monitoring Interviewers Using Paradata

Given the importance of reading the questions exactly as worded and the variability of interviewers' question-reading behavior, monitoring interviewers' behavior is arguably one of the more critical procedures for quality control processes. Monitoring interviewers' behavior for how they administer or read survey questions is done by listening to the interview recordings. Most survey organizations only listen to the first few recordings in their entirety or perform random spot checks for quality control, mostly due to resource limitations (Thissen & Myers, 2016; Viterna & Maynard, 2002). However, one could argue it is imperative to listen to later interviews, as research shows after an interviewer gains more experience with the survey, their interview lengths go down (Olson & Smyth, 2015; Couper & Kreuter, 2013). Some hypothesize the interviewers just become more familiar with the survey and their efficiency is increasing, but others argue this could be a sign they are breaking with standardized interviewing techniques

(Bradburn et al., 1979; Fowler & Mangione, 1990). Further, some surveys do not record interviews or record specific questions or sections, either because of the software or hardware limitations or because the interview contains sensitive questions. Because listening to interviews is resource-intensive and not always feasible, some organizations are using paradata, more specifically timestamps, as a proxy of how interviewers are administering questions.

Mick Couper originally conceptualized Paradata to describe the process data created as a by-product of computer-assisted data collection (Kreuter, 2013). However, Kreuter (2013) explains, since Couper's first inception of the term, paradata has expanded to include any "additional data that can be captured during the process of producing a survey statistics." Paradata can be captured either manually (e.g., interviewer records observations about their interactions with respondents) or automatically (e.g., the software captures when the interviewer presses a computer key to open a help screen). The types of paradata captured, whether manually or automatically, include interviewer call records, interviewer observations about fieldwork, keystroke data, and timestamps.

Using paradata to monitor interviewers is not a new concept. While most studies that use paradata to monitor interviewer behavior focus on contact rates, nonresponse, and sample assignment (Kirgis & Lepkowski, 2013; Wagner, 2013), some have used paradata to evaluate interviewer behavior during the interview. Keystroke paradata (e.g., pressing the F1 key, using the backspace key to back up in the interview) have been used as a proxy to identify issues with the survey instrument (Couper, 2000) and evaluation of interviewer performance on key performance indicators (Kirgis et al., 2015; Jans, Sirkis & Morgan, 2013).

More recently, organizations are using timing durations as proxies to indicate problems with how the interviewer reads the questions (Mneimneh, Pennell, Lin, & Kelley, 2014; Sun & Meng, 2014). Timestamps are created by the survey software, capturing the time from when the interviewer enters the screen (on which the question is displayed) until the point when the interviewer keys in the respondent's answer. Even though timing durations encompasses everything from the interviewer reading the question, to the respondent formulating and reporting their response, to the interviewer keying in the response, and possibly further interactions in between (e.g., probing, breaks away from interview), organizations use irregular timing durations as proxies to flag cases for further review. The theory is that too short timing durations may indicate the interviewer omitting words, paraphrasing, or skipping the question entirely, and too long timing durations may be an indicator of the interviewer adding words to the question.

The Saudi National Mental Health Survey used timing durations to flag questions read under one second to identify interviewers who may be skipping questions (Mneimneh, Pennell, Lin, & Kelley, 2014). The China Mental Health Survey, on a selected set of variables, used timing durations and minimum QATTs, calculated using the number of words in the question and reading pace (110 milliseconds per Chinese character), to flag suspect questions (Sun and Meng, 2014). While there are no known studies that use mean question reading times and standard deviations at the question-level to develop QATTs, studies do use mean section times (i.e., a set of questions), or overall interview lengths, to flag sections or interview lengths that fall outside a particular standard deviation (Mneimneh, Pennell, Lin, & Kelley, 2014; Murphy, Baxter,

Eyerman, Cunningham, & Kennet, 2004). The same process could be applied at the question-level to detect questions read outside an (x) standard deviation.

Using timing durations, along with QATTs, allows for more automated and targeted quality control. This is especially true for surveys that cannot record the whole interview or just parts of the interview. However, for surveys that record the entire interview, quality control efforts could be made more efficient. An automated flagging system could identify which questions violate established thresholds, and quality control staff could focus their efforts on flagged questions. Using question reading thresholds to monitor question reading times could also detect falsifying at the very first instance, or at the very least, detect interviewers who need more training in standard interviewing techniques.

However, the question remains about the accuracy of the detection methods mentioned above, which is best to detect question-reading deviations. Mneimneh et al. (2014) do not report the accuracy of flagging questions under one second. Sun and Meng (2014) reported the true deviation rate of the detection rate compared to the true deviation rate of randomly selected questions, but not the overall accuracy of their method. From their presentation, we can infer the rate of false-positive and verbatim, but to truly assess their detection method's accuracy, the rate of false-negatives (i.e., deviations the method is failing to identify) should be reported.

Sun and Meng (2014) use a WPS rate based on Chinese characters and cultural speech and comprehension rates to develop QATTs, which may differ from surveys conducted in English. Some organizations that conduct English language surveys instruct interviewers to read

questions at 2-3WPS, while others instruct interviewers to read at a normal conversation pace (Viterna & Maynard, 2002). Normal conversation rates can go as high as 250 words per minute (WPM) or 4.1 WPS (Foulke, E.. 1968), but listeners' comprehension starts to drop at 212 WPM or 3.5 WPS (Omoigui, He, Gupta, Grudin & Sanocki, 1999). Which WPS rate is best for detecting deviations for English-speaking interviewers? Given the variability of interviewers following study protocol on question-reading pace and the natural variability of speech rates in normal conversation, it is essential to test different WPS point-estimate rates for developing QATTs.

Also, Sun and Meng (2014) use a point-estimate (i.e., word count/110 millisecond per Chinese character) to flag any question's timestamp faster than the calculated WPS rate for a said question. However, the point-estimate is unidirectional and likely captures only deviations due to omitted words. One could argue that if one can estimate when a question is read 'too fast', one can estimate the point at which the question is read 'too slowly'. In theory, a WPS range could be used to create a minimum and maximum QATTs to flag both questions read 'too fast' (e.g., detecting omitted words) and questions read 'too slowly' (e.g., detecting added words), respectively.

Using a minimum WPS QATT makes theoretical sense; the minimum QATT is determined by estimating the minimum (i.e., fastest) time an interviewer can read the question without compromising the respondent's comprehension. As discussed previously, the questions flagged as 'too fast' may indicate interviewers reading faster than the prescribed reading pace or omitting

words or paraphrasing (i.e., a combination of omitting and substituting). In other words, minimum WPS QATTs do not have to factor in the respondent's behavior to detect this behavior.

Since the timestamp encompasses both interviewer and respondent behavior, a 'meaningful' maximum QATT should factor in the respondent's behavior. Research shows that respondents' response behavior (or process) is dependent on several factors, including the complexity of the question and respondents' cognitive abilities (Tourangeau, Rips, & Rasinski, 2000). The question's complexity does not necessarily increase as the number of words increases in the question; some 'short' questions can be just as cognitively challenging as 'long' questions. Thus, using the same maximum QATT for all questions, like the WPS range method does, may not be as accurate as methods that factor in respondents' behavior. Questions flagged by a WPS maximum QATT may be incorrectly flagging questions as question-reading deviations (i.e., false-positives), but the longer timestamp is due to the respondent's behavior (e.g., asked a question, thinking about the answer, or taking a break). However, the risk of increasing false-positives may be acceptable if the WPS range method detects more deviations than the WPS point-estimate method (or other methods). In other words, using a WPS range method may increase the number of deviations detected, but with an acceptable level of false-positives. Thus, the WPS Range method is worth investigating.

Considering the above discussion on WPS point-estimate and range methods, using measures of dispersion of data may be better at developing QATTS to detect deviations than using a WPS rate. For example, calculating QATTs using mean question-reading (i.e., timing durations) and standard deviations would allow interviewer speech rate variability and acknowledge the

timestamp also contains the respondent's response behavior. As stated previously, some organizations use mean interview time and standard deviations and flag interviews with suspicious lengths for further investigation. A natural inclination is to apply this to the question-level to detect suspicious question lengths.

However, this method may have its weaknesses. For one, it requires that sufficient data be available to reliably estimate a mean duration and standard deviation for each question. This method should be feasible early in data collection for longitudinal or cross-sectional surveys with paradata from previous waves. However, the method may not be informative for one-off surveys until enough data has been collected. The delay would most likely result in undesirable interviewer behavior not being corrected before the interviewer completed several interviews. Second, the behavior we want to detect (i.e., extreme question durations) influences means and standard deviations, thus influencing the QATTs. Nevertheless, to evaluate this method, several standard deviations (0.5; 1.0; 1.5; 2.0) will be tested in this study.

Finally, there may be a more accurate way of calculating QATTs than using WPS or standard deviations. One promising method borrows from a study (Munzert & Selb, 2015) that attempted to identify cheating in web surveys by modeling response latencies (i.e., timing durations from a web survey) as a function of person-specific random intercepts and fixed effects for the item (i.e., question) and whether or not the response was correct. Munzert & Selb (2015) then extracted the residuals from the model and categorized the top 2% observations as potential cheaters. They argued this analysis method isolated the "suspicious" response latency (at the question-level) from "latency that can be explained by systematic, as well as item- and person-

specific factors." This method could be used to develop QATTs; instead of using the timestamp as a proxy of response latency for web survey respondents, the timestamp is a proxy for interviewer question-reading times. However, Munzert & Selb (2015) did not discuss why they chose 2% as the thresholds, and it is not unreasonable to think that different thresholds may be better than others for trying to detect different behaviors. Thus, several percentage levels should be investigated to see which top and bottom percentage is most accurate for detecting question-reading deviations.

Research Questions

Given the importance of interviewers reading questions verbatim and the need to monitor their behavior, coupled with the growing use of paradata to increase the efficiency of quality control, both in terms of cost and time, the main research question is: Can timing durations be used as a proxy to detect interviewer question-reading deviations? More specifically, which of the above methods (i.e., WPS, standard deviation, and model-based) is best to establish QATTs for detecting question-reading deviations? Moreover, which rate or range should one use?

The following methods for developing QATTs (and the varying rates or range) will be compared on: 1) overall accuracy for correctly detecting questions read verbatim and questions read with deviations; 2) the proportion of correctly detected questions with major deviations; 3) proportion of correctly detected different types of deviations:

- WPS Point-Estimate Method
    - 2WPS; 3WPS; 4WPS
- WPS Range Method
    - 2WPS – 3WPS; 1WPS – 3WPS; 2WPS – 4WPS; 1WPS – 4WPS
- Standard Deviation Method
    - 0.5; 1.0; 1.5; 2.0
- Model-based Residual Method
    - 1%; 2%; 3%; 5%; 10%; 25%

1.3 Data and Methods

Sample

This study combines paradata and audio interview recordings from Wave 3 of the Understanding Society Innovation Panel. Understanding Society is a household panel study interviewing 40,000 households in the UK on various social and economic topics. The Innovation Panel (IP) is a separate panel for methodological research, with the results taken into consideration in the development of the next wave's main stage instruments (Killpack & Gatenby, 2010). The IP uses a multi-stage probability sample with an initial household CAPI interview to determine eligibility and collect household-level information. The target sample size for Wave 1 was 1500 households, and addresses were randomly selected from the Postcode Address File (PAF). Respondents who completed an interview at Wave 1 were invited to participate in subsequent waves. For Wave 3, 1526 eligible households were identified, and 1027 household interviews were completed with a response rate of 67%. The sample for Wave 3 was a mixture of both productive and unproductive Wave 2 households resulting in a response rate lower than expected (Killpack & Gatenby, 2010). All eligible adults (age 16+) in the household were then selected to complete an individual, face-to-face, computer-assisted personal interview (CAPI). Conditional on the household response rate, the individual response rate was 82%, for a total of 1621 completed interviews. The average interview length was 37.5 minutes, and interviewers are instructed to read all questions verbatim. Selected sections of the interview were recorded with the respondent's permission (72% consent rate, 1167 interviews). However, due to procedural and technical difficulties, only 820 interview recordings were available for analysis. The timing file contained timing durations for all interviews. However, specific questions that looped in the

questionnaire (i.e., the same question asked for different instances or situations) did not have a one-to-more match with the timing file. These questions were excluded from the analysis.

Behavior Coding

Behavior coding is widely used to study interviewer and respondent behavior in survey interviews by applying systematic coding to question-answer sequences (Cannell, Lawson, & Hausser, 1975; Ongena & Dijkstra, 2006a). As cited in Ongena and Dijkstra (2006a), Cannell Fowler and Marquis (1968) created the "first, fairly simple" coding scheme for interviewer-respondent behavior in surveys. Over the years, and as technology advanced, more sophisticated and complex coding was applied to datasets (Ongena and Dijkstra, 2006a).

This study's behavior coding builds on Cannell, Lawson, and Hausser's (1975) behavior coding scheme. Cannell, Lawson, and Hausser (1975) start with two broad codes: 1) "asks questions as printed" and 2) "asks question incorrectly." The behavior code then captures more detail for each; "Ask questions as printed" has two subcategories: 1) reading the question verbatim; and 2) "reads question making minor modifications of the printed version, but does not alter the frame of reference." "Asks questions incorrectly" has four subcategories that describe the type of deviation: 1) modifies or incorrectly reads response options; 2) significantly alters question (either main or stem); 3) does not read the question, but confirms anticipated response; and 4) "asks a question which should have been skipped."

These behavior codes were adapted for this study to create more refined subcategories for the type of deviation and described in detail below. Like Cannell, Lawson, and Hausser (1975), interviewers' first reading of the question was coded on whether or not they read the question

verbatim. If the interviewer did not read the question *exactly* as worded, it was coded as a deviation. The coding framework also included the type of deviation(s): omitted word(s) or substituted word(s), or added word(s). The categories are not mutually exclusive, and each question may have a combination of omitted, substituted, or added words. Like Cannell, Lawson, and Hausser's (1975), this study assumes deviations can impact the meaning of the question differentially; thus, deviations were then coded as minor and major deviations (see Figure 2.1). Minor deviations do not change the meaning of the question, and major deviations change the question's meaning.

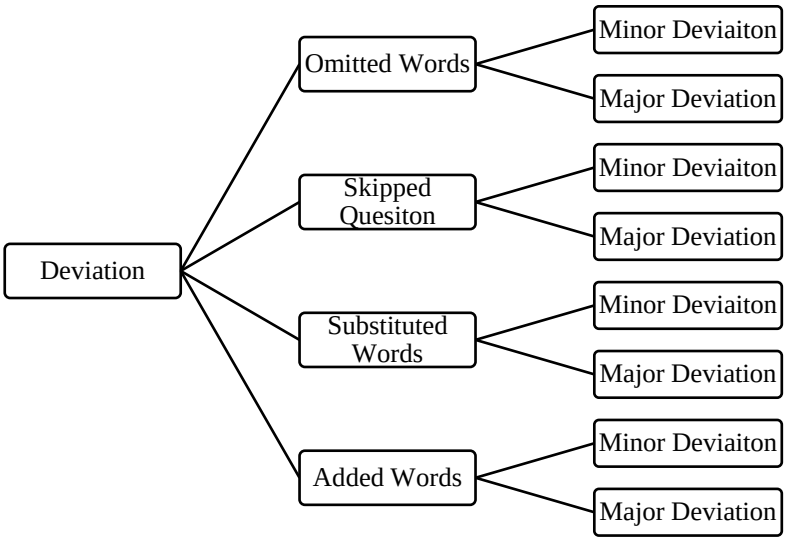

Figure 2.1 Behavior Coding for Deviations

Cannell, Lawson, and Hausser (1975) also give some guidance and examples on evaluating whether or not the deviation changed the meaning. Building on their definitions and examples, explicit rules (see Table 1.1) were created by this author to evaluate if the deviation was minor or major.

Table 1.1. Rules for Determining if Deviation was Minor or Major

Minor Deviations	Question as it Appears in Questionnaire	Examples of Deviations*
Omitted, subbed or added such words as "the", "a", "an" or words that did not give context to the question	The income of your household?	~~The~~ income of your household?
	Next, the income of your household, would you say you are dissatisfied, neither dissatisfied nor satisfied, or satisfied?	~~Next~~ *Now*, the income of your household, would you say you are dissatisfied, neither dissatisfied nor satisfied, or satisfied?
	Leaving aside your own personal intentions and circumstances, is your job a permanent job or is there some way that it is not permanent?	Leaving aside your own personal intentions and circumstances, is your job a permanent ~~job or is there some way that it is not permanent~~?
Interview instructions omitted, subbed or added that did not give meaning or context to question or to express politeness to the respondent	The next part of the survey is a little different. It has to do with memory and thinking.	The next part ~~of the survey~~ is a little different. It has to do with memory and thinking.
	Now, think of words that begin with the letter S as in Sarah. Start now.	Now, think of words that begin with the letter S as in Sarah. +Please+ Start now.
	People around here are willing to help their neighbours.	+Again using the showcard+ People around here are willing to help their neighbours.
Omitted a secondary time reference because secondary time reference was in previous question(s) or subbed a time reference that did not change reference period	The next questions ask about changes that may have happened to you since we last interviewed you on January 22, 2008.	The next questions ask about changes that may have happened to you since we ~~last~~ interviewed you on January 22, 2008.
	Since January 22, 2008, has a doctor or other health professional newly diagnosed you as having any of the conditions listed on the card? Please just tell me the numbers that apply.	Since ~~January 22, 2008~~ *your last interview*, has a doctor or other health professional newly diagnosed you as having any of the conditions listed on the card? Please just tell me the numbers that apply.
Interviewer omitted response options starting on the second question of a series of questions (e.g., always, very often, quite often, not very often, never) or respondent interrupted the interviewer to signal	(How often do you talk about politics or current affairs with...) Your (husband/wife/partner)? Always, very often, quite often, sometimes, rarely, never?	(How often do you talk about politics or current affairs with...) Your (husband/wife/partner)? Always, very often, quite often, sometimes, rarely, never?
	[*Next question*] (How often do you talk about politics or current affairs with...) Fellow workers? Always, very often, quite often, sometimes, rarely, never?	[*Next question*] ~~(How often do you talk about politics or current affairs with...)~~ Fellow workers? ~~Always, very often, quite often, sometimes, rarely, never?~~

their correct response for previously heard response options (e.g., agree, neither agree nor disagree, disagree)	People around here are willing to help their neighbours. Do you agree, neither agree nor disagree, or disagree? [*Next question*] People in this neighbourhood can be trusted. Do you agree, [respondent interrupts with "Disagree"] neither agree nor disagree, or disagree?	People around here are willing to help their neighbours. Do you agree, neither agree nor disagree, or disagree? [*Next question*] People in this neighbourhood can be trusted. Do you agree, [*Respondent interrupts with* "Disagree"] ~~neither agree nor disagree, or disagree?~~
Skipped the entire question, but response was given in previous answer	Do you [or anyone in your household] own a pet, such as a dog or cat? [*Next question*] What kind of pet do you own?	Do you [or anyone in your household] own a pet, such as a dog or cat? [*Respondent answers*, "Yes, we have a dog".] [*Next question*] ~~What kind of pet do you own?~~
Major Deviations	**Question as Appeared in Questionnaire**	**Examples**
Key nouns, verbs or adjectives/qualifiers were omitted	Do you have any store cards or credit cards such as Visa, or Mastercard in your sole name? Please do not include direct debit cards such as Switch or Delta or store loyalty cards such as Tesco Clubcard or Nectar.	Do you have any ~~store cards or~~ credit cards such as Visa, or Mastercard in your ~~sole~~ name? Please do not include direct debit cards ~~such as Switch or Delta or store loyalty cards such as Tesco Clubcard or Nectar.~~
	What is your current weight without clothes?	What is your current weight ~~without clothes~~?
	Do these health problem(s) or disability(ies) mean that you have substantial difficulties with any of these areas of your life? Please read out the numbers from the card next to the ones which apply to you.	Do these health problem(s) or disability(ies) mean that you have ~~substantial~~ difficulties with any of these areas of your life? Please read out the numbers from the card next to the ones which apply to you.
Key nouns, verbs or adjectives/qualifiers were subbed with words that did not have equivalence in meaning or were added that altered the context, added inaccurate meaning to the	I am going to read out a set of statements that could be true about your neighbourhood. For each, tell me whether you strongly agree, agree, neither agree nor disagree, disagree or strongly disagree that the statement describes your neighbourhood. First, this is a close-knit neighbourhood.	I am going to read out a set of statements that could be true about your neighbourhood. For each, tell me whether you ~~strongly~~ *agree*, *somewhat agree*, neither agree nor disagree, *somewhat disagree* or ~~strongly~~ *disagree* that the statement describes your neighbourhood. First, this is a close-knit neighbourhood.

question, or potential biased respondent's answer	In your household who has the final say in big financial decisions?	In your household who has the final say in big financial decisions? +Would you say you do?+
	And how do you usually get to your place of work?	And how do you usually get to your place of work? +Your car?+
Definitions or examples were omitted that were needed to give context to the question	About how often do you and people in your neighbourhood do favours for each other? By favours we mean such things as watching each other's children, helping with shopping, lending garden or house tools, and other small acts of kindness. Would you say often, sometimes, rarely or never?	About how often do you and people in your neighbourhood do favours for each other? ~~By favours we mean such things as watching each other's children, helping with shopping, lending garden or house tools, and other small acts of kindness.~~ Would you say often, sometimes, rarely or never?
	Do you save any amount of your income for example by putting something away now and then in a bank, building society, or Post Office account other than to meet regular bills? Please include share purchase schemes, ISA's and Tessa accounts.	Do you save any amount of your income ~~for example~~ by putting something away now and then in a bank, ~~building society, or Post Office account other than to meet regular bills? Please include share purchase schemes, ISA's and Tessa accounts.~~
Non-common response options were omitted that were needed to give context to the question to ensure all respondents were received same range of options	Do you work for a private firm or business or other limited company or do you work for some other type of organization?	Do you work for a private firm or business ~~or other limited company or do you work for some other type of organization?~~
Response options in a series of questions given for first time were omitted	On a scale from 1 to 7 where 1 means 'Completely dissatisfied' and 7 means 'Completely satisfied', how satisfied or dissatisfied are you with the following aspects of your current situation. First, your health.	On a scale from 1 to 7 ~~where 1 means 'Completely dissatisfied' and 7 means 'Completely satisfied'~~, how satisfied or dissatisfied are you with the following aspects of your current situation. First, your health.
Skipped the entire question	Would you say you disagree somewhat or disagree strongly?	[*Interviewer skips question without respondent indicating the strength of their disagreement in previous question*]
Strikethrough = omit word(s); +Plus signs+ = added word(s); *Asterisks* = subbed word(s)		

Behavior Coding Sample

Reviewing the literature on using behavior coding for question-level, there is no consistent sample strategy or sample size (Blair, 1980; Dijkstra, 2002; Holbrook, Cho, & Johnson, 2006; Jans, 2010; Lepkowski, Siu, & Fisher, 2000; Marquis & Cannell, 1969; Moore & Maynard, 2002; Ongena, 2005; Van der Zouwen & Dijkstra, 2002). Sample methods range from randomly selecting a subsample of interviews to selecting all interviews, arguably dependent on resources and funds. Sample size range from as few as 39 interviews (Moore & Maynard, 2002) to as many as 372 interviews (Blair, 1980). Total sample size range from 500 "verbal acts" (Marquis & Cannell, 1969) to 13,514 question administrations (Holbrook, Cho, & Johnson, 2006).

Given the above review and resource limitations, behavior coding was conducted on a subset of the available interview recordings (n=820). To select a subset of the recorded files for behavior coding, two interviews were randomly selected from each of the 80 interviewers. In a few cases, the selected interviews were missing recordings at the section level, resulting in only a few recorded questions in the interview. When this happened, an additional interview was randomly selected from the same interviewer to ensure that each interviewer had at least 50 questions coded[2]. This procedure yielded 168 interviews selected for behavior coding.

Within the selected interviews, 402 questions were selected for analysis based on the following criteria: Question was intended to be read out loud

- Did not contain 'fills'
- Were administered to both males and females

[2] This dataset is used in multiple studies, including examinations of question characteristics and interviewer effects. To increase analytic power, a minimum of 50 questions per interviewer was established.

- Had one-to-one matching with timing file questions (i.e., did not loop)
- Had the same response options for all regions

Due to question routing, not all questions were administered to all respondents. The total sample size for coding and analysis is 10,386 question administrations. The behavior coding was done directly from the audio files (no transcription) by a single coder. The behavior coded data is used as the 'gold standard' to which the deviation detection methods will be tested for accuracy.

Behavior Coding Variables

Using the behavior coding, a question-reading variable was created with three levels: 1) entire question read verbatim, 2) the question only contained minor deviation, and 3) the question contained at least one major deviation. Table 1.2 shows the distribution of the question-reading variable. Questions had minor-only deviations 34.5% of the time and major deviations 13.0%.

Table 1.1. Distribution of Question-Reading Variable (n=10386)

Question-Reading	Count	%
Verbatim	5447	52.5
Minor Deviation	3586	34.5
Major Deviation	1353	13.0

Additionally, a variable that describes the deviation type by magnitude was also created with the following levels: 1) Minor Omit Only, 2) Minor Substitute Only, 3) Minor Add Only, 4) Minor Multi Deviation, 5) Major Omit Only, 6) Major Substitute Only, 7) Major Add Only, 8) Major Multi Deviation, and 9) Verbatim. Of the 4939 deviations (both minor and major), almost three-

quarters (73.7%) of the deviations are due to interviewers only omitting words in questions (see Figure 2.2). Interviewers engage in only substituting words less often (11.4%) and even lesser, only adding words (4.7%) in questions. Interviewers make multiple types of deviations in a single question 10.2% of the time. Examining the different types of major deviations (see Figure 2.3), the majority is due to interviewers only omitting words in questions (84.6%) and rarely subbed (only) and added (only) words that changed the meaning of the question, 2.2%, and 1.4% respectively.

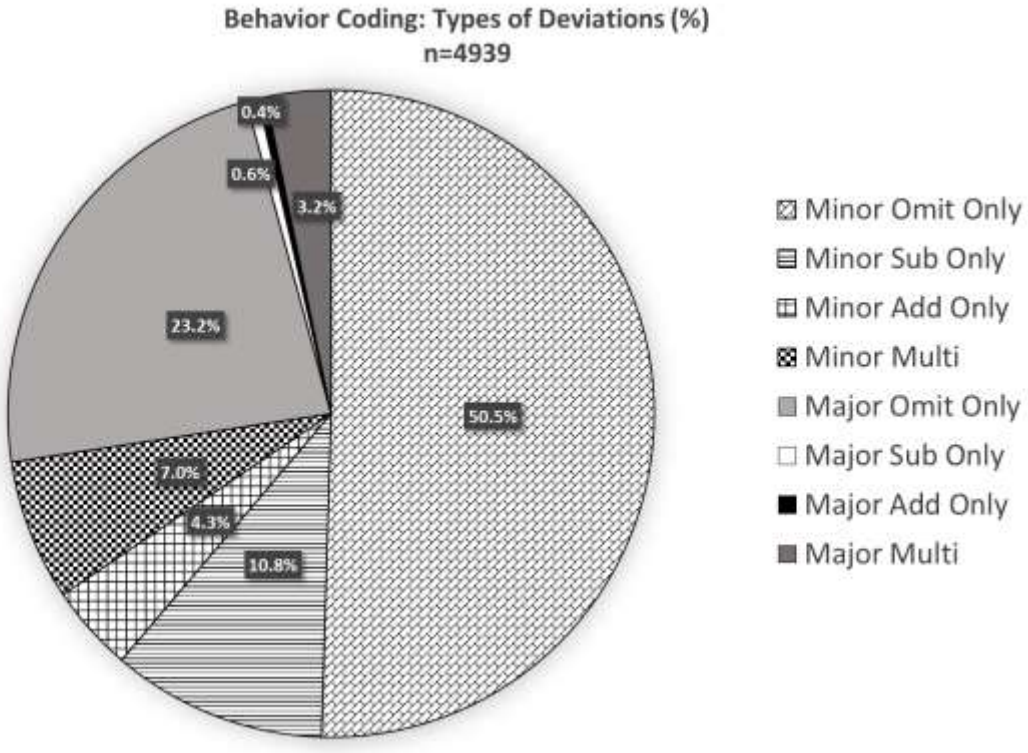

Figure 2.2. Distribution of Types of Deviations

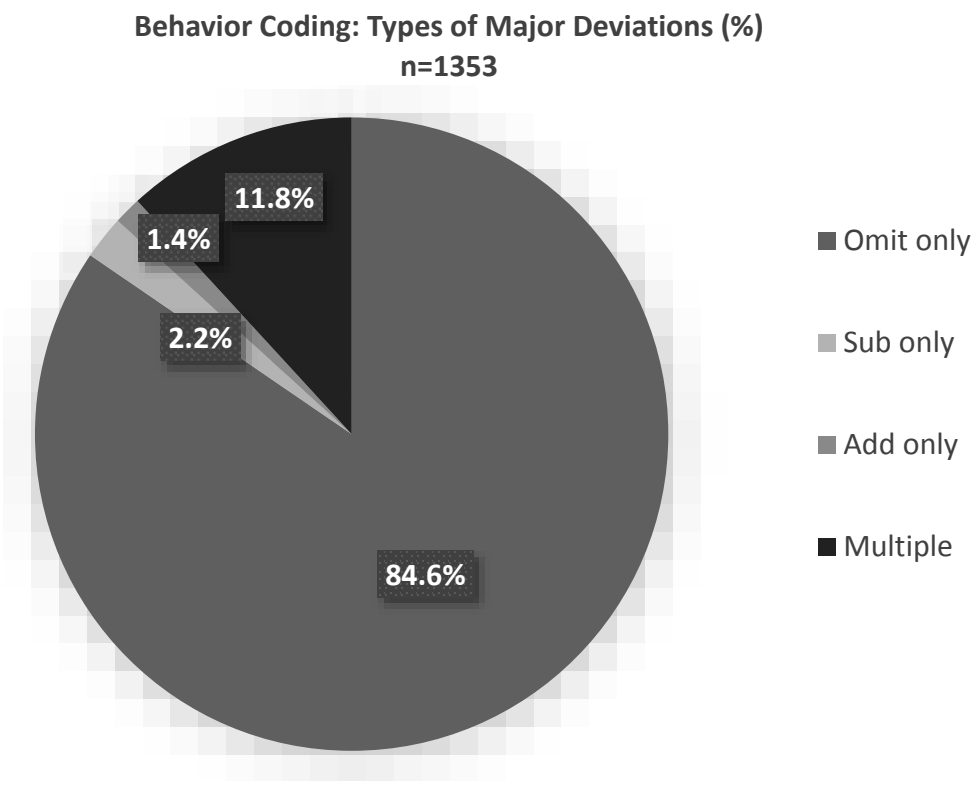

Figure 2.3. Distribution of Types of Major Deviations

QATT Detection Methods Variables

The next step was to create the detection method variables using the following QATT methods: 1) based on words per second; 2) using an 'x' standard deviation from the mean reading time of the question across interviewers and 3) using a model-based approach and classify the top and bottom (x) percentage of residuals as deviations. Below, each method is discussed in detail.

Words per Second Point-estimate and Range Methods

Using the words per second (WPS) point-estimate method flags questions that are read outside a certain threshold. As stated previously, interviewers are often instructed to read questions at a 2-3WPS pace to facilitate respondent comprehension (Viterna & Maynard, 2002). However,

conversation rates can go as high as 4.1, with comprehension starting to decrease at 3.5 WPS (Omoigui, He, Gupta, Grudin & Sanocki, 1999). However, the goal here is not to measure the interviewer's reading pace but to devise a systematic strategy for accurately detecting question-reading deviations. To that end, point-estimate thresholds were calculated at 2WPS, 3WPS, and 4WPS, by taking the total number of words (not including optional text) and dividing it by 2, 3, and 4, respectively. Any timestamp that was faster than (i.e., below) the point-estimate was flagged as a possible deviation. The following binary variables (0=no flag, 1=flagged for possible deviation) were created: 1) 2WPS; 2) 3WPS; and 3) 4WPS.

For the WPS range method, thresholds were also selected based on instructed pace and comprehension rates: 1) 2-3WPS and 2) 2-4WPS. Again, the total number of words (not including optional text) was divided by the WPS rate. However, the upper bound of the rate is the minimum QATT, and the lower bound is the maximum QATT. Using 2-3WPS as an example, time durations that were faster than (i.e., below) the 3WPS point-estimate were flagged as a possible deviation or any timestamp slower than (i.e., above) the 2WPS point-estimate was also flagged as a possible deviation. For example, a question with 12 words, the threshold for 2WPS is six seconds, for 3WPS the threshold is four seconds. For the 2-3WPS range, if the question duration is less than four seconds *or* more than 6 seconds, the question is flagged for either being too fast or too slow (respectively). The maximum QATT was also extended to 1WPS for each of these ranges to test an additional maximum QATT. The following binary variables (0=no flag, 1=flagged for possible deviation) were created: 1) 2-3WPS; 2) 2-4WPS; 3) 1-3WPS; and 4) 1-4WPS.

Standard Deviation Methods

The standard deviation method flags questions that are 'x' standard deviations from the mean question reading time across interviewers. Thresholds were calculated for this method by subtracting and adding 0.5, 1.0, 1.5, 2.0 standard deviations to the question mean, resulting in four detections methods: 1) 0.5 standard deviation (above and below); 2) 1.0 standard deviation (above and below); 3) 1.5 standard deviations (above and below) and 3) 2 standard deviations (above and below). Again, any timestamp below or above the (x) standard deviation was flagged as a possible deviation. The following binary variables (0=no flag, 1=flagged for possible deviation) were created: 1) SD 0.5; 2) SD 1.0; 3) SD 1.5; and 4) SD 2.0.

Model-Based Methods

As previously discussed, Munzert and Selb (2015) argue that response times (for web surveys) are a function of person-specific random effects and fixed effects for the question. Their multi-level model isolates suspicious response times from response times that can be explained by person-specific factors and the specific item (i.e., question) and whether or not the respondent had a correct answer (i.e., cheaters should take longer to answer). They then extracted the residuals and coded the top 2% as cheaters.

Extending this method to interviewer-administered questions, the question administration time (i.e., timing durations) is also likely to be specific to the respondent and question, but also to the interviewer and whether or not they read the question verbatim. Using the method of Munzert and Selb (2015) may isolate possible deviations from these factors.

Timing durations (logged) to each of the questions within interviewers are predicted by a model with a random intercept for the interviewer (Interviewer ID) and fixed effects for and the respondent (Respondent ID), each question (Question ID), and whether the question was read verbatim or not (0=Verbatim 1=Deviation). The interviewer random-effect variance estimate suggests there is some respondent (interviewer) level differences in question administration time (ICC = 0.164), and significant fixed effects were found for both the question and whether the question was read verbatim or not. The residuals from the model were standardized into a t-score to categorize the upper and lower (x) % of the t-distribution as possible deviations. As stated previously, Munzert & Selb (2015) do not discuss why they chose 2% as the threshold; thus, additional sets of upper and lower percentages were tested: 1%, 2%, 3%, 5%, 10%, and 25%. Table 1.3 shows the different percentage points and the upper and lower bound t-scores used to create the minimum and maximum QATTs for each following binary variables (0=no flag, 1=flagged for possible deviation): 1) Model 1%; 2) Model 2%; 3) Model 3%; 4) Model 5%; 5) Model 10%; and 6) Model 25%.

Table 1.2. Percentages and T-scores

Percentage	T-scores (lower, upper)
1%	-1.7978, 1.7315
2%	-1.5358, 1.4953
3%	-1.3623, 1.3498
5%	-1.1549, 1.1689
10%	-0.8547, 0.8969
25%	-0.4352, 0.4301

Table 1.4 shows the frequency (count and percentages) of *potential* deviations detected for each of the QATT detection methods. The other detection methods were parsed into deviations detected as ‘too fast’ and deviations detected as ‘too slow’ to make a fair comparison to the WPS

point-estimate method. For discussion purposes, all methods will be referred to by their variable names.

Reviewing the results for questions detected as ‘too fast’ first, the 2 WPS point-estimate method detected the highest rate of potential deviations (51.1%), followed by Model 25% (31.3%) and then SD 0.5 (28.3%). The 3WPS point-estimate and the WPS range methods that

Table 1.3. Potential Deviations Detected by QATT Detection Methods (n=10386).

	Detected 'Too fast' (Minimum QATT)		Detected 'Too Slow' Maximum QATT		Total Deviations Detected	
	Count	%	Count	%	Count	%
2WPS	5304	51.1	-	-	5304	51.1
3WPS	2347	22.6	-	-	2347	22.6
4WPS	1255	12.1	-	-	1255	12.1
2-3WPS	2347	22.6	4765	45.9	7112	68.5
1-3WPS	2347	22.6	1366	13.2	3713	35.8
2-4WPS	1255	12.1	4765	45.9	6020	58.0
1-4WPS	1255	12.1	1366	13.2	2621	25.2
SD 0.5	2927	28.2	2494	24.0	5421	52.2
SD 1.0	733	7.1	1675	16.1	2408	23.2
SD 1.5	145	1.4	1234	11.9	1379	13.3
SD 2.0	23	0.2	959	9.2	982	9.5
Model 1	397	3.8	456	4.4	853	8.2
Model 2	590	5.7	690	6.7	1280	12.4
Model 3	806	7.8	910	8.8	1716	16.6
Model 5	1127	10.9	1207	11.7	2334	22.6
Model 10	1797	17.4	1776	17.2	3573	34.5
Model 25	3236	31.3	3151	30.5	6387	61.7

include 3WPS point-estimate, detect the fourth-highest rate of potential ‘too fast’ deviations (22.6%). As the minimum QATTs become stricter for all methods, fewer potential ‘too fast’ deviations are being detected. For questions detected as ‘too slow’, the WPS range methods that

include the 2WPS point estimate (2-3WPS and 2-4WPS) detected the highest rate of potential deviations (45.9%), followed by Model 25% (31.3%). Like the minimum QATT, as the maximum QATTs become stricter for all methods, fewer potential 'too slow' deviations are being detected.

Combining the 'too fast' and the 'too slow' detected deviations, the 'least strict' version of each method detects higher rates of possible deviations within each method. The behavior coding identified 4393 (47.5%) deviations (both minor and major). The 2WPS point-estimate method is the closest to the behavior coding but overestimates the deviation rate, as does 2-3 WPS, 2-4WPS, SD 0.5, and Model 25% methods. The rate of false-positive and false-negatives for each of the methods is still unknown. A series of crosstabs will be performed to determine the accuracy of each detection method.

Analysis Methods to Determine Accuracy of QATT Detection Methods

Accuracy is defined as the rate of true-negatives and true-positives. It may be that some methods have high accuracy but are not useful because the method is failing to detect deviations (i.e., false-negatives) or creating too many red-herrings (i.e., false-positives). Thus the rate of false-negative, false-positive, true-negatives (i.e., verbatim), and true-positives (i.e., deviations) will be presented. First, the behavior coding variable was recoded as 0=Verbatim and 1=Any Deviation (i.e., combined minor and major). Then, a crosstab was performed for the behavior coding variable (0=Verbatim; 1=Any Deviation) and each QATT detection method to establish the rate of false-negatives (i.e., question deviations incorrectly identified as verbatim), false-positives (i.e., verbatim questions incorrectly detected as deviations), true-negative (i.e., verbatim question correctly identified), and true-positive (i.e., question deviations correctly

identified). This analysis will produce the accuracy rate for detecting any deviation, as well as the utility of each of the methods. As with other studies using a 'gold-standard' comparison to survey data (e.g. Davern et al. 2008; Goldman, Chu, Osmond, and Bindman, 2011; Short et al., 2009; Tang, Ralston, Arrigotti, Qureshi, and Graham, 2007), the percent concordant is used to identify overall accuracy.

To assess the accuracy for detecting major deviations, arguably what is of most interest, the behavior coding variable was recoded as 0=Verbatim/Minor Deviation and 1=Major Deviation. Crosstabs were performed with each of the methods to establish false-negatives, false-positives, true-negatives, true-positives. Next, crosstabs were performed using these new variables with the type and magnitude variable (0=Verbatim, 1=Minor Omit Only, 2=Minor Substitute Only, 3=Minor Add Only, 4=Minor Multi Deviation, 5=Major Omit Only, 6=Major Substitute Only, 7=Major Add Only, and 8=Major Multi Deviation) to evaluate if some QATT detection methods are better for detecting certain types of deviations. The results are reported and discussed in the next section.

1.4 Results

Table 1.5 shows the rate of false-negatives, false-positives, true-negatives, and true-positives for each QATT detection method by Any Deviation and also parsed into "too fast" and "too slow" detections. By adding the rate of true-negative and true-positive, we can determine each method's overall accuracy for detecting 'too fast' deviations, 'too slow' deviations' and total deviations. Looking first at the total deviations detected, the QATT method having the highest accuracy for detecting any deviation is 3WPS, 67.2%, and the least overall accurate method is 1-3WPS (45.2%), with the remainder of the methods falling in between 49.4% and 65.9%.

Table 1.4. Accuracy Rate (%) of Detecting Deviations: QATT Detection Methods by Any Deviation (n=10386)

	Detected 'Too fast' (Minimum QATT)					Detected 'Too slow' (Maximum QATT)					Total Deviations Detected				
	False Neg	False Pos	True Neg	True Pos	Overall Acc	False Neg	False Pos	True Neg	True Pos	Overall Acc	False Neg	False Pos	True Neg	True Pos	Overall Acc
2WPS	15.3	18.8	33.7	32.3	65.9	-	-	-	-	-	15.3	18.8	33.7	32.27	65.9
3WPS	28.8	3.9	48.5	18.7	67.2	-	-	-	-	-	28.8	3.9	48.5	18.69	67.2
4WPS	36.2	0.8	51.7	11.3	63.0	-	-	-	-	-	36.2	0.8	51.7	11.33	63.0
2-3WPS	28.9	3.9	48.5	18.7	67.2	33.5	31.9	20.6	14.0	34.6	14.8	35.8	16.7	32.71	49.4
1-3WPS	28.9	3.9	48.5	18.7	67.2	43.7	9.3	43.1	3.9	47.0	22.2	32.6	19.8	25.35	45.2
2-4WPS	36.2	0.8	51.7	11.3	63.0	33.5	31.9	20.6	14.0	34.6	25.0	13.2	39.2	22.54	61.8
1-4WPS	36.2	0.8	51.7	11.3	63.0	43.7	9.3	43.1	3.9	47.0	32.4	10.1	42.4	15.18	57.6
SD 0.5	29.7	10.3	42.2	17.9	60.0	37.7	14.2	38.2	9.8	48.1	19.9	24.5	28.0	27.70	55.7
SD 1.0	42.0	1.5	51.0	5.6	56.6	40.7	9.3	43.2	6.9	50.1	35.1	10.7	41.7	12.46	54.2
SD 1.5	46.3	0.1	52.3	1.3	53.6	42.4	6.7	45.8	5.2	51.0	41.1	6.8	45.6	6.47	52.1
SD 2.0	47.3	52.4	0.0	0.2	0.2	43.4	5.1	47.4	4.1	51.5	43.2	5.1	47.4	4.37	51.7
Model 1	44.4	0.8	51.8	3.1	54.7	45.4	2.4	50.2	2.0	52.0	42.3	3.1	49.4	5.11	54.3
Model 2	43.1	1.4	51.2	4.3	55.3	44.5	3.7	48.8	3.0	51.6	40.1	5.1	47.5	7.32	54.6
Model 3	41.7	2.1	50.5	5.7	56.0	43.5	4.8	47.7	4.0	51.5	37.8	6.9	45.6	9.70	55.1
Model 5	39.8	3.2	49.3	7.6	56.7	42.2	6.4	46.2	5.3	51.3	34.5	9.6	42.9	12.95	55.7
Model 10	36.2	6.1	46.4	11.2	57.4	40.0	9.7	42.8	7.4	50.0	28.8	15.9	36.7	18.66	55.1
Model 25	29.4	13.2	39.3	18.1	57.2	34.4	17.4	35.1	13.1	48.0	16.3	30.6	21.9	32.09	53.8

The ‘net’ can become too big or too small for overall accuracy as it converges on the true-negative rate (i.e., verbatim) and the true-positive rate (i.e., deviations). For example, the 2WPS method overestimates (i.e., the ‘net’ is too big) the true deviations (for any deviations) by 7.4%, and accuracy starts to decline due to the increase of false-positives. If the ‘net’ is made smaller (i.e., increasing the WPS pace) than 3WPS, overall accuracy starts to decline due to false-negatives. Even though the table does not display it for all methods, this reasoning extends to the other methods; when the method overestimates deviations, overall accuracy decreases, but the rate of detecting true-negatives (i.e., deviations) increases.

Looking at the accuracy for detecting questions ‘too fast’ and ‘too’ slow’ can better understand how minimum and maximum rates might mitigate the overall accuracy rate. For example, the 3WPS, 2-3WPS, and 1-3WPS methods have the highest rate of overall accuracy for detecting ‘too fast’ deviations at 67.2%. However, when examining the accuracy rates for maximum QATTs, both ranges have relatively lower accuracy rates for detecting ‘too slow’ deviations, and thus the overall accuracy rate for detecting any deviation decreases. Using the SD 2.0 method as an example, the overall accuracy rate of 51.7% is mostly due to the method’s ability to detect ‘too slow’ deviations; the accuracy rate for SD 2.0 for detecting ‘too fast’ is negligible (0.2%). The WPS methods have higher rates of detecting questions read ‘too-fast’, while the standard deviation and model-based methods have higher rates of detecting ‘too-slow’. These results suggest using a combination of methods may increase overall accuracy.

The argument could be made that merely due to human-error, all interviews will contain some deviations. Coupled with the argument that minor deviations do not change the meaning of the question, the focus should be on detecting major deviations best. Table 1.6 shows the accuracy

Table 1.5. Accuracy Rate (%) of Detecting Deviations: QATT Detection Methods by Major Deviation (n=10386)

	Detected 'Too fast' (Minimum QATT)					Detected 'Too slow' (Maximum QATT)					Total Deviations Detected				
	False Neg	False Pos	True Neg	True Pos	Overall Acc	False Neg	False Pos	True Neg	True Pos	Overall Acc	False Neg	False Pos	True Neg	True Pos	Overall Acc
2WPS	2.6	40.6	46.4	10.5	56.8	-	-	-	-	-	2.6	40.6	46.4	10.46	56.8
3WPS	4.9	14.5	72.5	8.1	80.6	-	-	-	-	-	4.9	14.5	72.5	8.14	80.6
4WPS	6.9	6.0	81.0	6.1	87.1	-	-	-	-	-	6.9	6.0	81.0	6.10	87.1
2-3WPS	4.9	14.5	72.5	8.1	80.6	10.6	43.5	43.5	2.4	45.9	2.5	57.9	29.0	10.55	39.6
1-3WPS	4.9	14.5	72.5	8.1	80.6	12.3	12.4	74.5	0.7	75.2	4.5	49.4	37.5	8.52	46.1
2-4WPS	6.9	6.0	81.0	6.1	87.1	10.6	43.5	43.5	2.4	45.9	4.2	26.9	60.1	8.85	68.9
1-4WPS	6.9	6.0	81.0	6.1	87.1	12.3	12.4	74.5	0.7	75.2	6.2	18.4	68.6	6.82	75.4
SD 0.5	6.0	21.1	65.8	7.0	72.9	11.0	21.9	65.0	2.1	67.1	3.9	43.1	43.9	9.11	53.0
SD 1.0	9.8	3.8	83.2	3.2	86.4	11.4	14.5	72.4	1.6	74.0	8.2	18.4	68.6	4.82	73.4
SD 1.5	12.1	0.4	86.5	1.0	87.5	11.7	10.6	76.4	1.3	77.7	10.8	11.0	75.9	2.15	78.1
SD 2.0	12.8	0.0	86.9	0.2	87.1	12.0	8.2	78.8	1.0	79.8	11.8	8.3	78.7	1.19	79.9
Model 1	11.6	2.5	84.5	1.3	85.8	12.6	4.1	83.0	0.3	83.3	11.3	6.6	80.5	1.68	81.8
Model 2	43.1	1.4	51.2	4.3	55.5	44.5	3.7	48.8	3.0	51.8	10.7	10.1	76.9	2.28	78.9
Model 3	10.9	5.7	81.3	2.1	83.5	12.2	8.0	79.0	0.8	79.9	10.0	13.6	73.4	2.94	76.0
Model 5	10.3	8.2	78.8	2.7	81.4	11.8	10.5	76.5	1.2	77.7	9.1	18.7	68.3	3.86	71.9
Model 10	9.4	13.8	73.2	3.6	76.8	11.2	15.4	71.6	1.8	73.4	28.8	15.9	36.7	18.66	55.1
Model 25	7.5	25.8	61.2	5.5	66.7	9.8	27.3	59.8	3.2	63.0	16.3	30.6	21.9	32.09	53.8

results for major deviations. Looking at overall accuracy for total major deviations detected first, the highest overall accurate QATT method is 4WPS, 87.1%, followed by the Model 1% method (81.8%). The least overall accurate method is 2-3WPS (39.6%), with the remainder of the methods falling in between 46.1% and 80.6%. Examining accuracy for minimum QATTs for detecting major deviations that are 'too fast', the highest overall accurate method is SD 1.5 (87.5%), but it is just 0.4 percentage points above 4WPS point estimate (also, 2-4WPS and 1-4WPS) and the SD 2.0 method. Of those methods, the 4WPS (also, 2-4WPS and 1-4WPS) has the lowest rate of false-negatives (i.e., deviations not detected), but it has a higher rate of false-positives than the standard deviation methods.

Looking at accuracy for maximum QATTs (i.e., 'too slow') for detecting any deviation, the Model 1% method has the highest rate of accuracy at 83.3%, and the least accurate method is the 2-3WPS at 45.9%. Again, these results suggest a hybrid of methods may increase overall accuracy. However, is overall accuracy the goal? A particular method may have high rates for overall accuracy, but the high rate is due to accurately identifying true-negatives (i.e., verbatim) and only detects little or no deviations.

Table 1.7 displays the rate of detecting any deviations and major deviations (i.e., true-positives/ [false-negatives + true-positives]) for each detection method. For any deviations, as stated previously, the method that has the highest overall accuracy for detecting any deviations is 3WPS (67.2%). However, the detecting rate for detecting true deviations (any) is only 39.4%. Six other methods detected more true deviations than 3WPS, with 2WPS detecting the most true deviations at 67.9%. The same holds for major deviations; 4WPS has the highest accuracy rate for detecting any major deviations (87.1%), but the rate for detecting major deviations is only

46.9%. Six other methods detect more major deviations, with the 2-3WPS method detecting the most major deviations (81.0%). However, increasing the rate of detecting deviations comes at a price; the number of false-positives can increase dramatically. The 2-3WPS detects 81% of the major deviations, but the false-positive rate soars to 57.9%. This is because minor deviations are classified as false-positives, but the methods cannot differentiate between minor and major deviations. This may be acceptable if the goal is to target major deviations only. However, quality control staff would spend a significant amount of time chasing down and ruling out red-herrings (i.e., false-positives).

Table 1.6. Detection Rate (%) of Any Deviations Detecting and Major Deviations Detected by Methods

	Any Deviation		Major Deviations	
	% Deviations Detected n=4939	Overall Accuracy	% Deviations Detected n=1353	Overall Accuracy
2WPS	67.9	65.9	80.3	56.8
3WPS	39.4	67.2	62.5	80.6
4WPS	23.8	63.0	46.9	87.1
2-3WPS	68.8	49.4	81.0	39.6
1-3WPS	53.3	45.2	65.4	46.1
2-4WPS	47.4	61.8	67.9	68.9
1-4WPS	31.9	57.6	52.3	75.4
SD 0.5	58.3	55.7	69.9	53.0
SD 1.0	26.2	54.2	37.0	73.4
SD 1.5	13.6	52.1	16.6	78.1
SD 2.0	9.2	51.7	9.2	79.9
Model 1%	10.8	54.3	13.0	81.8
Model 2%	15.4	54.6	17.6	78.9
Model 3%	20.4	55.1	22.6	76.0
Model 5%	27.3	55.7	29.7	71.9
Model 10%	39.3	55.1	41.4	55.1
Model 25%	66.3	53.8	66.9	53.8

However, false-positive and false-negatives may be reduced if the data is aggregated up to the interview level. In quality control, when questions are flagged as suspicious by other quality

control procedures (e.g., too many questions entered as don't know or refused, the backup key used too often, outside the expected range), it is illogical to think only those questions flagged are investigated. In most cases, the activity leading up to the suspicious question(s) and the subsequent behavior is assessed, and in some cases, the entire interview is reviewed. If an interview has questions flagged as having potential question-reading deviations, listening to the interview should catch the deviations that the QATT method missed (i.e., false-negatives) and rule out the deviations the method identified as verbatim (i.e., false-positives).

To that end, the data were aggregated to the interview level. The SAS procedure 'proc tabulate', for the interview number and a variable indicating whether the question was categorized as a false-negative, false-positive, true-negative, or true-positive, was used to create a new dataset at the interviewer level for each detection method, for both any deviation and major deviations. The new datasets contained the interview number (rows; n=168) and four variables (columns): count of false-negative, count of false-positive, count of true-negative, and count of true-positive. From this, two new variables were created: 1) Interview has true deviation [if (false-negative or true-negative) > 0, then interviewer has true deviation=1, else=0]; and 2) Method detected deviation [if (false-positive or true-negative) > 0, then method detected deviation=1, else=0].

The last step was creating four variables: 1) Method correctly identified the interview contains deviations (0=no, 1=yes); 2) Method correctly identified the interview contains no deviations (0=no, 1=yes); 3) Method *incorrectly* identified the interview contains deviations (0=no, 1=yes); and 4) Method *incorrectly* identified the interview contains no deviations (0=no, 1=yes). Frequencies were run for each of the variables for each of the detection method. Accuracy rates were derived for the totals and reported, along with the detecting rate for detecting true

interviews with at least one major deviation. Finally, the rate of interviews the method flagged for further investigation was also calculated (i.e., correctly identified as containing major deviation(s), plus incorrectly flagged as having deviations/total interviews).

First, all interviews contained at least one minor deviation, and all but 29 interviews contained at least one major deviation. Almost 83% of the interviews would require further investigation makes the following discussion somewhat moot. However, only 168 interviews were behavior coded. It could be that a larger dataset or a different subsample of the interview recordings would have produced fewer interviews with major deviations. We can learn from this analysis if we focus on the accuracy rate and the detection rate for correctly identifying the 139 interviews with major deviation(s). Also, one could argue that at the start of field operations, the first interviews completed by each interview may have a high rate of interviews with major deviations. Thus, ruling out 17% of the incoming interviews for needing review would reduce quality control efforts. For this reason, the results are displayed and discussed for major deviations (see Table 1.8).

This analysis aims to see if any of the methods correctly detected 139 interviews containing major deviations and correctly identified 29 interviews as verbatim (i.e., containing no major deviations). Unfortunately, none of the methods reaches 100% overall accuracy. The accuracy ranges from 79.2% (Model 1%) to 88.7% (4WPS). Similar to the question level analysis, having a high accuracy rate does not mean the method is best at identifying interviews that have major deviations; the methods with the lower accuracy rates detect higher rates of true interviews with major deviations but also flags almost all, if not all interviews as needing further review.

Table 1.7 Interview Level Analysis for Ruling Out False-positives and Discovering False-negatives (n=168)

Detection Method	Count of Interviews Correctly Flagged As Containing:		Count of Interviews Incorrectly Flagged as Containing:		Overall Accuracy (%)	% of Interviews Deviation Detected n=139	Interviews Method Flagged for Review (%)
	Deviation	**No Deviation**	**Deviation**	**No Deviation**			
2WPS	139	0	29	0	82.7	100.0	100.0
3WPS	137	6	23	2	85.1	98.6	95.2
4WPS	132	17	7	12	88.7	95.0	82.7
2-3WPS	139	0	29	0	82.7	100.0	100.0
1-3WPS	139	0	29	0	82.7	100.0	100.0
2-4WPS	139	0	29	0	82.7	100.0	100.0
1-4WPS	138	4	25	1	84.5	99.3	97.0
SD 0.5	139	0	29	0	82.7	100.0	100.0
SD 1.0	139	3	26	0	84.5	100.0	98.2
SD 1.5	134	10	19	5	85.7	96.4	91.1
SD 2.0	124	13	16	15	81.5	89.2	83.3
Model 1	127	6	23	12	79.2	91.4	89.3
Model 2	133	2	27	6	80.4	95.7	95.2
Model 3	137	2	27	2	82.7	98.6	97.6
Model 5	139	1	28	0	83.3	100.0	99.4
Model 10	139	0	29	0	82.7	100.0	100.0
Model 25	139	0	29	0	82.7	100.0	100.0

If an organization's goal is to detect all interviews with any major deviation, no matter the increase of false-positives, out of all the methods that correctly identify 100% of the 139 interviewer containing major deviations, Model 5 is the only one to correctly rule out one interview. If the goal is to reduce quality control efforts while acknowledging that some interviews that contain major deviations may not be detected and some may be missed, then the 4WPS may be the best method; 17 (10.1%) interviews can be ruled out as needing further review and the method only incorrectly identifies seven (4.2%) interviews as containing deviations (i.e., red herrings) and 12 (7.1 %) interviews as having no deviations.

Turning to the last research question; is one QATT detection method better than another method for detecting the different types of deviations? Given the results from the previous analysis, it is no surprise the 2WPS method has the highest rate for detecting deviations due to words only being omitted (see Table 1.9), for both minor (70.4%) and minor (83.7%). For deviations due to interviewers substituting words only, for minor deviations, Model 25% has the highest detection rate (66.9%), but the 2-3WPS rate detects the highest rate for major deviations (63.3%). For deviations due to interviewers adding words only, for minor deviations, Model 25% has the highest detection rate (62.1%), and again, the 2-3WPS rate detects the highest rate for major deviations (78.9%). For deviations due to interviewers making multiple types of deviations, for minor deviations, Model 25% has the highest detection rate for both minor (71.6%) and major (72.5%). However, we know from previous discussions that although one method may be better at detecting different types of deviations, it does not mean it is necessarily the best method to use. As detection rates increase, so does the rate of false-positives.

Table 1.8. Detection Rate of Different Types of Deviations by Methods

	Detection Methods	Omitted Words	Substituted Words	Added Words	Multiple Types
	Behavior Coding n	2493	531	214	348
Minor Deviations	2WPS	70.4	49.2	24.8	56.3
	3WPS	36.6	16.9	4.7	23.9
	4WPS	19.4	3.8	0.0	11.2
	2-3WPS	63.1	63.1	75.7	66.1
	1-3WPS	42.5	31.3	34.6	35.1
	2-4WPS	45.9	49.9	71.0	53.4
	1-4WPS	25.3	18.1	29.9	22.4
	SD 0.5	56.7	44.1	46.7	52.6
	SD 1.0	22.1	19.8	24.3	24.7
	SD 1.5	11.3	14.3	15.0	14.4
	SD 2.0	8.5	11.1	11.7	9.5
	Model 1	8.7	10.9	10.3	16.4
	Model 2	13.2	15.4	14.0	23.0

	Model 3	18.1	21.3	19.6	27.0
	Model 5	24.4	27.9	29.4	35.1
	Model 10	36.3	40.5	41.6	47.7
	Model 25	63.5	66.9	62.1	71.6
	Behavior Coding n	1144	30	19	160
	2WPS	83.7	56.7	26.3	66.3
	3WPS	68.0	20.0	5.3	37.5
	4WPS	51.6	10.0	5.3	25.0
	2-3WPS	83.1	63.3	78.9	69.4
	1-3WPS	72.0	36.7	42.1	47.5
	2-4WPS	66.7	53.3	78.9	56.9
	1-4WPS	55.6	26.7	42.1	35.0
Major Deviations	SD 0.5	72.4	53.3	47.4	58.1
	SD 1.0	38.7	20.0	31.6	28.8
	SD 1.5	17.7	10.0	26.3	13.8
	SD 2.0	9.2	3.3	21.1	8.8
	Model 1	13.5	3.3	5.3	11.3
	Model 2	17.7	6.7	10.5	18.1
	Model 3	22.9	6.7	21.1	22.5
	Model 5	29.2	13.3	31.6	34.4
	Model 10	41.2	20.0	36.8	45.0
	Model 25	66.0	46.7	73.7	72.5

1.5 Conclusions

Words matter. Especially in survey research. Researchers know changes in question-wording can change the meaning of the question, thus changing the question's validity (Groves et al., 2011; Krosnick, Malhotra, & Mittal, 2014; Schuman & Presser, 1996). For this reason, interviewers are trained and instructed to read questions exactly as worded. However, like previous research (Ackermann-Piek and Massing, 2014; Cannell, Lawson, & Huasser, 1975; Mathiowetz & Cannell, 1980), this study finds that interviewers engaged in question reading deviations almost half of the time (47.5%) they read a question to the respondent. The majority (73.7%) of the deviations interviewers made for this sample was omitting words from the question text.

Deviations were mostly minor, meaning they did not change the meaning of the questions, but almost 26% of the deviations committed resulted in changing the question's meaning. Hence, giving further proof that monitoring interviewers' question-reading behavior could affect data quality.

One way to monitor interviewer question-reading behavior is by listening to audio interview recordings. However, this work is resource-intensive, and some surveys cannot be recorded. Given that paradata can be collected with relative ease and with little cost using survey software, some organizations have started using paradata, more specifically timing durations, as a proxy of how interviewers are administering questions to flag suspect timing durations at the question-level. However, the utility for using timing durations in this manner is unknown. To flag suspect questions, organizations must first develop a question-administration timing threshold (QATT) and then compare them to the questions' timing duration. There is no established or tested way to develop QATTs.

This study tested a known method (i.e., WPS point-estimate method) and three methods not previously used to develop QATTs (i.e., WPS range method, standard deviations of mean question-reading times, and model-based). To assess the accuracy and the utility, each QATT method was compared to the behavior coded data (i.e., used as the 'gold standard' for this study). Results show that the most overall accurate QATT method for detecting any *potential* deviation is the 3WPS (67.2%). However, one could argue that the goal is not to find the most overall accurate method for developing QATTs to detect question-reading deviations but to find the QATT method for best detecting deviations. Further, since interviewers are human and 'to err is human', it is reasonable to assume most, if not all interviews, will contain at least minor

deviations, and the effort should be focused on detecting major deviations. The results show that all 168 interviews contain at least one minor deviation, and 139 interviews contain at least one major deviation.

For major deviations, the method with the highest overall accuracy rate is 4WPS (87.1%), but the 2WPS method is best at detecting potential major deviations (80.3%). Along with failing to detect actual deviations (i.e., false-negatives), the 2WPS produces the highest rates of false-positives. So the utility of using 2WPS comes into question. One might think that aggregating the data up to the interview level might reduce false-positives and false-negatives for the 2WPS method. Some utility is gained with aggregating the data up to the interview level, but not for the 2WPS method; the 2WPS method does correctly identify all interviews containing at least one major deviation, it incorrectly identifies 29 interviews has having deviations; thus, 100% of all interviews are flagged for further review. The method that arguably shows the most utility at the interview level is 4WPS; it has the highest rate of correctly identifying interviews with no major deviations (10.1%), while only incorrectly identifying 4.2% of interviews containing deviations, and 7.1% of interviews as having no deviations. This targeted, automated approach should save time and money by reducing the need to listen to all interviews and concentrating quality control efforts on those interviews (or interviewers) with high rates of questions (or interviews) flagged as having major deviations.

Whereas these methods show considerable promise in this study, there is still a significant amount of research that can be done in this area. One easily identified area is to assess how (or if) taking question-reading deviations affects data quality. We assume minor deviations do not impact data quality, but major ones do impact data quality. More research is needed to

understand how question-reading deviations affect data quality. Developing QATTs for surveys conducted in different languages is another area of research that has not been explored. Would 4WPS still show the most promise for other languages as it does for English?

Limitations

This study is the first known to show that survey paradata, which is relatively inexpensive to collect, can be used to develop QATTs that can identify major question misreadings with reasonable success. This method has considerable potential to improve the efficiency of field monitoring. However, the study does have limitations. First, while the behavior coding was a unique feature of the data that allowed the study to be conducted, it was only performed on a subset of the interview recordings due to technical, administrative, and resource limits. While random sampling should ensure that the coded interviews are a representative subsample of all recorded interviews, there is a risk that the interviews that were not recorded differ from those that were recorded. Interviewers who engage in more question-reading deviations may not want to be recorded and may take steps to manufacture a 'technical' issue (e.g., unplugged or turned off the microphone) or falsely indicate that the respondent refused to be recorded. Thus, having a more complete sample or a different sample may change the results. Second, even with a carefully developed coding scheme and coding criteria, behavior coding as a method does involve some subjectivity. Even with these limitations, this study suggests that establishing and using QATTs is a promising method to improve quality control processes in interviewer-administered surveys and is deserving of additional research.

Question Characteristics and Interviewer Question-Reading Deviations

Abstract

When interviewers deviate from script (i.e., omit, substitute or add words) they may be changing the meaning of the question and thus impacting measurement error. Given the importance of reading questions exactly as worded and the numerous studies that report question-reading deviations, there are only a handful of studies that attempt to identify the cause of why interviewers engage in this behavior; behavior that has the potential to negatively impact data quality. This study focuses on the impact question characteristics have on question-reading deviations in face-to-face interviews. Specifically, are there certain types of questions that have a higher probability of interviewers making question-reading deviations? Using behavior-coded data from Wave 3 of the Understanding Society Innovation Panel, this study investigates which question characteristics (e.g., type of question, length, complexity, etc.) lead to an increase in question-reading deviations that have a high probability of changing the meaning of the question

2.1 Introduction

A well-known tenet in survey question design is to keep it 'short and simple'. The main objective of this tenet is to improve question comprehension and reduce respondent cognitive burden. However, translating this (i.e., drafting short and simple questions) into practice is often challenging, especially when the question's intent is to measure a complex behavior or attitude. Also, deciding on which question structures or characteristics to use (e.g., giving an example or definition, providing a showcard, number, and type of response options) is often challenging with conflicting guidelines. Thus, surveys often include long and complex questions, and questionnaire designers often use different question characteristics to ask the same question.

In standardized interviewer-administered surveys, the interviewer is tasked with reading all questions exactly as worded, including long and complex questions. However, research has

shown interviewers often go off script (Ackermann-Piek & Massing, 2014; Belli & Lepkowski, 1996; Cannell, Lawson, & Huasser, 1975; Haan, Ongena & Huiskes, 2013; Mathiowetz & Cannell, 1980; Oksenberg, Cannell & Kalton, 1991). When interviewers deviate from the script (i.e., omit, substitute, or add words), they may be changing the meaning of the question and thus impacting measurement error. (Groves et al., 2011; Krosnick, Malhotra, & Mittal, 2014; Rugg, 1941; Schuman & Presser, 1996). For example, if interviewers do not read "without clothes" when asking the question "What is your current weight without clothes?" the respondents' answer will most likely differ than if the interviewer did include "without clothes".

Given the importance of reading questions precisely as worded and the numerous studies that report question-reading deviations, only a handful of studies attempt to identify the cause of why interviewers engage in this behavior, behavior that has the potential to impact data quality negatively. Schober and Conrad (2002) hypothesize that interviewers may go off script because they are trying to help the respondent comprehend a question the interviewer perceives as a "bad" question. Others argue that interviewers tailor the question to the respondent, letting the respondent know they are listening and incorporating previously volunteered information into the interviewer-respondent interaction (Haan, Ongena & Huiskes, 2013; Ongena & Dijkstra, 2006b).

Another reason may be due to lack of training or experience. Interviewers may shorten or skip questions to speed up an interview with an uncooperative respondent (e.g., the respondent is displaying survey fatigue or irritation). Other reasons may be less noble; interviewers may

intentionally deviate from script to shorten the interview for personal gain (e.g., they are paid by the interview). The question then becomes, why do interviewers deviate from script?

This study focuses on the impact question characteristics have on question-reading deviations in face-to-face interviews. Specifically, are there certain types of questions that have a higher probability of interviewers making question-reading deviations? This study also begins to explore the impact of respondent and interviewer characteristics on question-reading deviations. Using behavior-coded data from Wave 3 of the Understanding Society Innovation Panel, this study investigates which question characteristics (e.g., type of question, length, complexity) increase interviewers' odds of engaging in major deviations.

2.2 Background

Question Characteristics and Interviewer Question-Reading Behavior Literature

The literature on question characteristics and interviewers' question-reading behavior is sparse and only examine a few question characteristics at broad levels, such as open-ended questions versus closed-ended questions. Studies that examined question-reading deviations and open-/closed-ended questions report conflicting results; three studies found that open-ended questions were less likely to be read verbatim than closed-ended questions (Bradburn, Sudman, Blair, Locander, Miles, Singer & Stocking, 1979; Cannell & Robison, 1971; Mathiowetz & Cannell, 1980), but Cannell, Miller, and Oksenberg (1981) found the opposite. Bradburn et al. (1979) also examined question length and found shorter questions were less likely to have deviations than longer questions containing extraneous introductions and more formal language.

Presser and Zhao's 1992 study extended the above-cited research by adding additional question characteristics. The study coded 94 survey questions on four question characteristics: Length (number of words in the question); Position (where it appears in the survey); Familiarity (the proportion of times the question was asked over the course of the study); Series (wording is almost identical to the previous question). The study also examined interviewer characteristics, experience, refusal rate, and efficiency. The results show that interviewers made more deviations as the question length increases and when the question was part of a series. Position and familiarity, and interviewer characteristics were not associated with how the interviewer read the question. The authors conclude, like previous studies (and textbooks), questions should be short and add "…brevity helps interviewers do their jobs as well" (p. 239).

While the authors offer the above guidelines, "short" and "brevity" are vague. Measuring when a question meets these guidelines is difficult. Also, there may be other question characteristics contributing to how interviewers read questions. Analyzing additional question characteristics should give questionnaire designers more detailed guidance for designing questions where standardized interviewing is the goal.

The Presser and Zhao study used data from a telephone lab, which have been found to have fewer interviewer deviations than face-to-face interviews (Ackermann-Piek and Massing, 2014; Bradburn et al., 1979; Cannell, Lawson, & Huasser, 1975; Cannell & Robison, 1971; Mathiowetz & Cannell, 1980; Cannell, Miller, and Oksenberg, 1981; Presser & Zhao, 1992). Lower deviation rates in telephone labs may be due to the fact that interviewers may be able to focus more easily on the screen and the question wording than a face-to-face interviewer. In

addition, interviewers in a face-to-face interview setting have the additional tasks of maintaining eye contact and keeping the respondent engaged and look for any non-verbal signs of confusion, fatigue, or distraction, and thus look away from the screen more so than in a telephone lab setting, thus resulting in more deviations.

However, the proximity of other interviewers and supervisors in centralized telephone labs could also be the reason for lower rates of deviations. In many cases, including telephone and face-to-face interviews, interviewers are routinely checked, either by supervisor observation or recordings, to see if they are reading questions verbatim as part of their performance evaluation. In centralized telephone facilities, observing and recording interviews can be done more easily, given the proximity of supervisors and the technology available.

In face-to-face surveys, supervision is done in the field. Sending a supervisor to the field to observe interviews is costly. For one day of the supervisor's time, they might observe three face-to-face interviews, compared to eight or more telephone interviews if in a central facility. The technology infrastructure in a telephone lab is often more sophisticated and can handle large audio files, and does not have to rely on broadband or cellular networks to transmit data, unlike the laptops used in face-to-face interviews. Because of the technology limitations, face-to-face interviews are less likely to be recorded, or only portions of the interview are recorded than telephone surveys.

One could argue that in a face-to-face interview setting where there are no (or few) supervisors or coworkers to observe their behavior and no (or few) recordings, interviewers may make more

deviations than interviews conducted in a lab. In other words, interviewers in the lab have more incentive to be on their 'best' behavior where just the proximity of other interviewers or coworkers or their supervisor can easily overhear whether or not they are following protocols. This pressure to conform to protocols is removed in face-to-face interviews.

In addition to making human-errors in question reading, face-to-face interviewers could intentionally change question-wording in unobserved and unrecorded interviews with little or no repercussions. Interviewers may realize this and feel embolden to engage in this behavior, especially if this behavior benefits the interviewer (e.g., completes more interviews). Hence, face-to-face interviewers may engage in more major question-reading deviations than telephone interviewers. Thus, question design guidelines for minimizing question-reading deviations resulting from telephone studies may not apply to face-to-face interviews as telephone studies may not be capturing the type of deviations face-to-face interviewers are making. There is a clear literature gap for a more detailed analysis of question characteristics and question-reading deviations for face-to-face interviews.

One limitation of studies is the narrow scope of the variables used for analysis (Bradburn et al., 1979; Cannell & Robison, 1971; Mathiowetz & Cannell, 1980; Cannell, Miller, and Oksenberg.1981; Presser & Zhao 1992). Including additional question characteristics variables should provide researchers with a better understanding of what impacts interviewers' question-reading. Past studies most likely were constrained by the difficulty of coding lengthy questionnaires on multiple characteristics, software limitations, and computing power regarding how many variables (and the level of variables) they could include in their analysis. However, as

computers become more powerful and efficient, and software becomes more sophisticated, a more comprehensive analysis is feasible. Further, the above studies' analysis did not differentiate between minor and major deviations and used multiple regression methods to examine the relationships. Given the dataset's hierarchical structure, multi-level modeling would be a more prudent analysis method (Gelman and Hill, 2006).

While there is a dearth of literature on question characteristics and interviewer question-reading behavior, there is quite an accumulation of research on question characteristics and respondent behavior. This literature focuses on how question characteristics have the potential to influence the different stages of the response process. While this research focuses on the respondent, these studies provide useful methodology to researchers investigating question characteristics and interviewer behavior, given the asking of questions is an interaction between the interviewer and the respondent.

Question Characteristics

This study categorizes question characteristics into three areas: 1) Structure, 2) Content, and 3) Question Aids. The following section discusses how certain elements of these areas might affect interviewers' question-reading behavior and interviewers' inclination to make major question-reading deviations. It should be noted, the question examples throughout this section are used to discuss each question characteristic separately. However, as stated previously, a question can have many characteristics; there may be other question characteristics that may be impacting question reading than the one being discussed.

Question Structures

Question structures can be thought of as *how* the question is designed. A gate question is one such type of question structure. Gate questions are defined as questions that, if answered in a certain way (most common "yes"), is followed by subsequent questions on the same topic. As interviewers (and some respondents) gain more experience with the survey, they are more likely to know which questions when answered a certain way, will make the interview longer (Couper & Kreuter, 2013; Olson & Peytchev, 2007). If for whatever reason (e.g., fatigued respondent or personal gain), interviewers want to shorten the question, interviewers may deviate from the script, so the question is posed in a way that the follow up questions are not triggered, including not reading the question aloud and enter 'no' to the gate question (Eckman et al., 2014). Thus, gate questions may be more likely to have major question-reading deviations than gate-dependent questions or independent questions (i.e., neither gate nor gate-dependent).

Another question structure that may impact interviewer question-reading is if the question is part of a series. A series question is defined as questions that appear one after another on the same topic where the response options are the same. An example is:

- How often do you and people in this neighbourhood have parties or get-togethers where other people in the neighbourhood are present? Would you say often, sometimes, rarely never?
- How often do you and other people in this neighbourhood visit each other's homes or chat to each other on the street? Would you say often, sometimes, rarely never?
- About how often do you and people in your neighbourhood do favours for each other? By favours we mean such things as watching each other's children, helping with shopping, lending garden or house tools, and other small acts of kindness. Would you say often, sometimes, rarely never?

As previously discussed, Presser and Zhao (1992) found that interviewers made more deviations when the question was part of a series. Interviewers who want to speed up the interview may decide to shorten questions as they go through the series as they may judge (either correctly or incorrectly) the respondent understands the questions are related and have the same response options. Interviewers may feel it is acceptable to shorten questions that are part of a series more so than questions that are not part of series, not merely because the question is longer. Including question length as a variable (and holding it constant) in the analysis allows us to examine the effect of being part of a series has on question-reading deviation regardless of question length.

Similar to series questions are questions that have a common stem. Common stem questions are part of series, but they have the same leading or ending text (i.e., question stems). The example given for part of series, also has a common stem (i.e., How often do you and people in this neighbourhood). Interviewers may feel like they can deviate from the script because only the question's subject is changing, while the majority of the text and response options stay the same. For example, omitting the common text, "How common in your area" in the first question will change the meaning of the question, as the deviation changes the question to a Yes/No question: "Is Rubbish or liter lying around?" However, if the first question is read verbatim, omitting the common stems in the subsequent questions will not change the questions' meaning. For example, if the interviewer only reads "Drunks or tramps on the street?", one may argue the respondent most likely understands the questions are asking how common these things are and have the same response options. So, although the stems are meant to be read, and omitting the stem is a deviation, the deviation is not a major deviation. Thus, overall questions with common stems should have a lower probability of major deviations as the structure, and repetitive nature of the

questions allows the interviewer to shorten the questions without changing the meaning (with the exception being the first question in the series) more so than questions that are not a part of a series. In other words, non-series questions cannot 'borrow' meaning from previous questions to retain their meaning when words are omitted. Thus they are more likely to have major deviations.

How response options are presented may also affect interviewers' likelihood to engage in major deviations. Questions that have response options read in the text of the question will have more words than questions that do not include the response options in the questions' text. Interviewers looking to shorten or speed up the interviewer process may not read all the response options. One of the tenets of standardized interviewing is that all respondents receive the same response options for a question. Not doing so may comprise the validity (i.e., meaning) of the question. Questions that include the response options in the text should have a higher probability of major question-reading deviations.

The type of response options is another question structure that previous research shows is associated with question-reading deviations (Bradburn et al., 1979; Cannell & Robison, 1971; Mathiowetz & Cannell, 1980; Oksenberg, 1981). However, the studies do not differentiate between minor and major deviation, only use the broad category of open or closed-ended questions, and show conflicting results, as stated previously.

Another question structure that might affect question reading is when the question is asked (i.e., the question order). It is widely thought that interviews speed up as the interview progresses. For

example, respondents may speed up the response process as they learn how to be a 'good' respondent, and in turn, the interviewer administers the questions quicker, or the interviewer speeds up in response to the respondent displaying signs of fatigue. However, one study found that interviewers took longer to administer questions as the interview progressed (Couper & Kreuter, 2013). The authors acknowledge their findings are unexpected, and they hypothesize the questions towards the end of the survey may be more difficult or burdensome, thus accounting for the longer times. However, a measure of question difficulty was not included as a covariate, and the authors acknowledge that future research should include one to explore this finding further, which this study does.

Question Content

Question content can be thought of as what the question is asking or what is in the question. Question length and difficulty are almost always included in question characteristics studies. Question length is often defined as the number of words in the question text the interviewer must read to the respondent (Bradburn et al., 1979; Presser and Zhao, 1992). Not only do more words in a question mean there is more opportunity for the interviewer to omit or substitute words, but more words most likely add to the complexity of the question. Thus, longer questions should be more likely to have question-reading deviations as some interviewers may want to shorten the interview or think they need to 'help' the respondents by simplifying the question. Indeed, past research has shown that longer questions are more likely to have deviations than shorter questions (Bradburn et al., 1979; Presser and Zhao, 1992)

As for question difficulty, interviewers may deviate from script to 'help' the respondent with question comprehension but unwittingly change the question's meaning. Question difficulty has been operationalized in many different ways. For example, Mangione, Fowler, and Louis (1992) categorize questions as either difficult or not difficult, and coded questions that "required the respondent to recall things that may be hard to remember….or dealt with an issue that was complicated or the respondent was unlikely to have thought much about before". Given that these factors most likely differ from respondent to respondent and requires a subjective determination (e.g., hard is a vague qualifier), other studies have used more objective measures of question difficulty such as question reading levels (Holbrook, Cho, & Johnson, 2006; Olson & Smyth, 2015). Using an objective difficulty measure should reduce any error that may be introduced by subjective coding. There are two measures of readability, Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKG). Velez and Ashworth (2007) argue that FKG is the more appropriate readability formula to use as it performs better with shorter pieces of writing, like survey questions, by using "the number of syllables per word regardless of the number of words" thus, this study uses FKG as an approximate of question difficulty.

Question content is not limited to question length and difficulty. Other elements of question content that have yet to be studied in relation to question-reading deviations include the type of question (e.g., demographic, attitudinal, behavioral, instructional), the number of response options, questions that confirm past wave information, double-barreled questions, and the sensitivity of the question. Olson and Smyth (2015) found quicker response times for attitudinal and demographic questions than behavioral, contradicting previous research that shows attitudinal questions take longer to answer (Bassili & Fletcher, 1991; Tourangeau, Rips, &

Rasinski, 2000; Yan & Tourangeau, 2008). Olson and Smyth (2015) argue the attitudinal questions were such that the respondent would have a ready answer, and conversely, the behavioral questions require more comprehension and retrieval effort. However, the quicker times for attitudinal and demographic questions could be attributed to the interviewer question-reading deviations (e.g., omit words or paraphrase), not necessarily that the respondent is answering attitudinal questions more quickly than other types of questions. The conflicting findings and not knowing whom to attribute the quicker response times (or longer response times) to, this study makes no predictions about which type of question is more likely to have question reading deviations.

Question sensitivity is another question characteristic that has not been studied in relation to question-reading deviations but has been studied in other survey research areas. Sensitive questions are questions where the respondent may edit their response due to embarrassment (if they answer a certain way) or to hide information from third parties (Tourangeau et al. 2000). Extending this sense of embarrassment to interviewers, interviewers may feel embarrassed to administer these questions or feel that the questions are too intrusive and hence edit the question (i.e., omit or change the wording or skip the question). Sensitive questions should have a higher probability of question-reading deviations.

Questions that confirm past wave information may also be more prone to deviations. Longitudinal studies can use previous wave information to be used in current wave questions. Past wave questions may be more familiar to interviewers (for those who were staffed for past waves). Other interviewers may try to 'help' respondents remember their previous responses in

relation to their current status. These deviations have good intentions, but they may change the meaning of the question.

Double-barreled questions have also yet to be studied in relation to question-reading deviations. Double-barreled questions are defined as containing more than one reference item, where the items could produce differing responses, but only one response option is offered. An example of a double-barreled question is, “How often do you talk about politics or current affairs with family members?”. Respondents could talk about current affairs (e.g., new stories or celebrities) but never talk about politics with families. With multiple items in the question, interviewers who want to speed up the interview may drop one of the items. Dropping one of the ‘barrels' most likely changes the meaning of the question. If the interviewer has experienced problems (e.g., the respondent does not know how to answer because only part of the question applies to their situation) with this question in previous interviews, the interviewer may alter the text to 'help' identify the question's intent. However, interviewers may be less likely to perceive that they can omit words without changing the meaning if the two barrels are distinct in meaning.

Question Aids

Question aids are another area that is under-studied in question-reading deviations research. Questionnaire designers use question aids for many reasons, including trying to make the interviewer process more efficient and to aid the interviewer in administering the question and the respondent in answering the question. Question aids vary from showcards (i.e., cards given to the respondent that have the response options listed) to optional text or definitions and examples

to help text (information available to the interviewer if the respondent is having difficulty giving a codable response).

Aids, such as help text or optional text, are thought to help the interviewer process, but one could argue that interviewers may feel like they can be less exact with the question wording knowing there is 'backup' help available and may be more susceptible to question-reading deviations. For questions with definitions or examples, interviewers may think some respondents *need* the definition or example or time reference, while other more capable (e.g., more educated) respondents do not need the definition or example. Some questions are also structured so that the example or definition is optional (by placing in parentheses), so the interviewer may think that omitting an example or definition is acceptable because they are not always mandatory to read. Interviewers may think they are maintaining the intent of the question, but without the context of the definition or the example, they may be changing the meaning of the question.

Time references also provide crucial cues to the respondent to facilitate retrieval (Tourangeau, Rips, and Rasinski (2000). Examples of time references are "Since we last interviewed you…" or "In the last <time boundary>, have you…". However, the further the interview progresses, the interviewer may feel like it is unnecessary to read time references, especially repeated references, and therefore questions with time references may be more prone to deviations than questions without time references. A study done by Uhrig and Sala (2011) found that interviewers failed to read time boundaries 33% of the time.

Showcards provide the respondent with a visual aid for answering questions. Interviewers may feel they have more leeway in how they read the question knowing the respondent has a showcard to refer to when answering the question. This study predicts that questions with these characteristics (i.e., aids) should have a higher probability of deviations.

Respondent and Interviewer Characteristics and Question-Reading Deviation

While this study's focus is question characteristics, respondent and interviewer characteristics will be used as control variables in this study. It would be remiss not to discuss the literature for respondent and interviewer characteristics impact on question-reading deviations. Again, the literature is sparse. Bradburn et al. (1979) found that older and more experienced interviewers make more deviations than younger and less experienced, but the differences were not statistically significant. Presser and Zhao (1992) found that interviewer experience, refusal rate, and efficiency were not significantly correlated with interviewer question-reading deviations. However, as stated previously, these interviews were conducted in a telephone lab, not in a face-to-face setting where interviewers' behavior may be different. This study will fill a gap in the literature by evaluating interviewer characteristics and question-reading deviations in a face-to-face context and adding additional interviewer characteristics.

There are no known studies of respondent characteristics and interviewer question-reading deviations. However, we know from response time research that some respondent characteristics, such as age and cognitive ability, have been found to contribute to longer response times (Couper & Kreuter, 2013; Yan & Tourangeau, 2008). One explanation for this is that interviewers may pick up on this and consequently add words or phrases to 'help' respondents. However,

interviewers may recognize the need to slow down and carefully read the question to older and less cognitively able respondents and feel they can speed up the process with younger and more cognitive-able respondents. Including these variables in the analysis should give further insight. This study will also include other respondent characteristics as controls that may have an impact on question-reading.

2.3 Data and Methods

Sample

This study combines paradata derived from audio interview recordings (i.e., interview behavior coded data) and the questionnaire (i.e., question characteristic coded data) from Wave 3 of the Understanding Society Innovation Panel. Understanding Society is a household panel study interviewing 40,000 households in the U.K. on various social and economic topics. The Innovation Panel (IP) is a separate panel for methodological research (i.e., experiments and testing questions, procedures, and methods in a context similar to the main study) with the results taken into consideration in the development of the next wave's main stage instruments (Jäckle, Gaia, Al Baghal, Burton & Lynn, 2017). The I.P. uses a multi-stage probability sample with an initial household CAPI interview to determine eligibility and collect household-level information. The target sample size for Wave 1 was 1500 households, and addresses were randomly selected from the Postcode Address File (PAF). Interviews were conducted at 1489 households (59.0% response rate), and 2393 individual interviews were completed, with an 88.9% conditional individual response rate. Respondents who completed an interview at Wave 1 were invited to participate in subsequent waves. For Wave 3, 1525 eligible households were identified, and 1027 household interviews were completed with a response rate of 73.9%. All

eligible adults (age 16+) in the household were then selected to complete an individual, face-to-face, computer-assisted personal interview (CAPI). Conditional on the household response rate, the individual response rate was 82.2%, for a total of 1621 completed interviews. The average interview length was 37.5 minutes, and interviewers are instructed to read all questions verbatim. Selected sections of the interview were recorded with the respondent's permission (72% consent rate). However, due to procedural and technical difficulties, only 820 interview recordings were available for analysis. The timing file contained timestamps for all interviews. Certain questions that looped in the questionnaire (i.e., same question asked for different instances or situations) did not have a one-to-one match with the timing file. These questions were excluded from the analysis.

Behavior Coding

Interview recordings were behavior-coded into three categories: 1) questions asked verbatim, 2) those with minor deviations, and 3) those with major deviations. Behavior coding has a long history in studying interviewer behaviors (Cannell, Fowler & Marquis, 1968; Cannell, Lawson, & Hausser, 1975; Dijkstra, 2002; Fowler & Cannell, 1996; Houtkoop-Steenstra, 2000; Ongena & Dijkstra, 2006a). Some studies simply code interviewer question-reading behavior as verbatim or not verbatim (Mangione, Fowler, & Louis, 1992; Peneff, J. 1988), while others code the degree of the deviation (Belli & Lepkowski, 1996; Oksenberg, Cannell, & Kalton, 1991). However for those studies who code the degree of deviations, they often do not define or give examples of what constitutes 'change the meaning' (Belli & Lepkowski, 1996; Oksenberg, Cannell, & Kalton, 1991) and those that do operationalize the coding for 'change the meaning' do so in varying degrees of specificity Cannell, Lawson, & Hausser, 1975).

One study that gives some guidance and examples on determining if deviations change the meaning of the questions is the widely cited work by Cannell, Lawson, and Hausser (1975). Building on the authors' definition and examples, this study constructed an explicit set of rules to determine if the question was read verbatim and had minor deviations or major deviations. Table 2.1 shows some examples of the coding rules for major deviations (see Chapter 1 for detailed coding).

Table 2.9. Sample of Behavior Coding Rules

Major Deviations	**Question as Appeared in Questionnaire**	**Examples**
Key nouns, verbs or adjectives/qualifiers were omitted	Do you have any store cards or credit cards such as Visa, or Mastercard in your sole name? Please do not include direct debit cards such as Switch or Delta or store loyalty cards such as Tesco Clubcard or Nectar.	Do you have any ~~store cards or~~ credit cards such as Visa, or Mastercard in your ~~sole~~ name? Please do not include direct debit cards ~~such as Switch or Delta or store loyalty cards such as Tesco Clubcard or Nectar.~~
	What is your current weight without clothes?	What is your current weight ~~without clothes~~?
Non-common response options were omitted that were needed to give context to the question to ensure all respondents were received same range of options	Do you work for a private firm or business or other limited company or do you work for some other type of organization?	Do you work for a private firm or business ~~or other limited company or do you work for some other type of organization?~~
Response options in a series of questions given for first time were omitted	On a scale from 1 to 7 where 1 means 'Completely dissatisfied' and 7 means 'Completely satisfied', how satisfied or dissatisfied are you with the following aspects of your current situation. First, your health.	On a scale from 1 to 7 ~~where 1 means 'Completely dissatisfied' and 7 means 'Completely satisfied'~~, how satisfied or dissatisfied are you with the following aspects of your current situation. First, your health.
Skipped the entire question	Would you say you disagree somewhat or disagree strongly?	[*Interviewer skips question without respondent indicating the strength of their disagreement in previous question and enters what they think is the response the respondent would give.*]

Key nouns, verbs or adjectives/qualifiers were subbed with words that did not have equivalence in meaning or were added that altered the context, added inaccurate meaning to the question, or potential biased respondent's answer	In your household who has the final say in big financial decisions?	In your household who has the final say in big financial decisions? +Would you say you do?+
	And how do you usually get to your place of work?	And how do you usually get to your place of work? +Your car?+
Strikethrough = omit word(s); +Plus signs+ = added word(s); *Asterisks* = subbed word(s)		

Behavior Coding Sample

Studies that use behavior coding vary in their sample methods. Given the breadth of resources, studies have used sample sizes from 39 interviews to 372 interviews and varying sample strategies from selecting a subsample of interviews to selecting all interviews and likewise for question selection (Blair, 1980; Dijkstra, 2002; Holbrook, Cho, & Johnson, 2006; Jans, 2010; Lepkowski, Siu, & Fisher, 2000; Marquis & Cannell, 1969; Moore & Maynard, 2002; Ongena, 2005; Van der Zouwen & Dijkstra, 2002). This study selected a subset of the available interview recordings (n=820). The recordings were stratified by interviewer (n=80), and two interviews were randomly selected from each interviewer. Additional interviews were randomly selected from interviewers where the two coded interviews did not result in at least 50 coded questions. This method would ensure that every interviewer was represented in the data. In all, 168 recorded interviews were selected for behavior coding.

Questions from the Wave 3 I.P. Questionnaire were included in the dataset on the following criteria if the question: 1) was intended for the interviewer to read the question aloud to the respondent; 2) did not have a varying number of words based on the previous answer or

respondent characteristics (i.e., fills), 3) were administered to both males and females (e.g., omitted fertility questions); 4) had a one-to-one match with timing file questions (i.e., did not loop); 5) had the same response options for all regions (i.e., did not include questions that have regional based response options); and 6) the question was recorded. The questions selected for analysis (n=361) were coded for each of the recordings sampled. Because routing through the questionnaire is dependent on respondents' answers, not all questions are asked of respondents. In total, 10,345 questions administrations were behavior coded for analysis.

Behavior Coding Variable

Using the above-mentioned behavior coding, a question-reading variable was created with three levels: 1) entire question read verbatim; 2) question only contained minor deviation; and 3) question contained at least one major deviation. This study focuses on the relationship between major deviations and question characteristics; therefore, verbatim and minor deviations were collapsed, resulting in a binary variable, 1) verbatim or minor-only deviations and 2) major deviations. The behavior coding shows that interviewers engaged in major question-reading deviations for 13% of the cases in this sample.

Question Characteristic Coding Variables

Question coding has been longed used to study both interviewer and respondent behavior and to evaluate question design. Like behavior coding, the dimensions coded and the operationalization of question coding varies from study to study. For example, Mangione, Fowler, and Louis (1992) code questions on four dimensions: Sensitive/Not sensitive; Difficult/Not Difficult;

Opinion/Factual; Open/Closed. Presser and Zhao (1992) code questions on four dimensions: Length, Position, Familiarity, and Series. An example of differing operationalization of question coding within dimensions, Mangione, Fowler, and Louis (1992), categorizes questions as open or closed questions, while Olson and Smyth (2015) code questions as open-ended text, open-ended numerical, closed nominal, closed-ordinal, yes/no.

To expand the literature on question characteristics and interviewer question-reading deviations, questions are expanded to 17 dimensions discussed in the Background section. Table 2.2 shows the distribution of the question characteristics. The first column shows the dimension and the levels coded. The second column shows the number of each question character for questions in the Wave 3 I.P. Questionnaire. The third column shows the percentage (or mean for continuous variables) coded for each question characteristic used in the analysis, and the fourth column is the standard deviation for the continuous variables.

Table 2.2. Descriptive Statistics for Question Characteristics

Question Characteristic	**Selected Questions n=361**	**%/mean**	**S.D.**
Structure			
Gate or Independent Question			
Gate	71	25.3	
Follow up Question	164	24.7	
Independent	126	74.7	
Part of Series of Questions	144	38.3	
Stem	60	18.7	
Response Options Read in Text	140	32.9	
Type of Response			
Yes/No	61	24.0	
Select one	185	45.8	
Select all that apply	19	4.3	
Scale	26	6.5	

Open-ended	57	10.1	
Content			
Word Count	361	25.8	19.7
Difficulty (FKG)	361	8.0	4.3
Type of Question			
Demo/Factual	73	17.6	
Behavioral	151	41.6	
Attitudinal/Belief	124	31.4	
Intro/Instruction	13	9.4	
Number of Response Options		3.6	3.3
Confirming Past Wave Information	25	10.5	
Double Barreled	24	9.2	
Sensitive Question	60	18.9	
Question Aids			
Optional Text	39	11.1	
Definition or Example Given	21	12.8	
Time Reference	20	9.1	
Showcard	96	23.0	
Question Help	41	10.9	

Respondent, Interviewer, and Interview Context Level Control Variables

This study will fill a gap in the literature by evaluating interviewer and respondent characteristics and question-reading deviations in a face-to-face context. Table 2.3 shows the distribution of the respondent, interviewer, and interview context variables used as controls in the models. The first column shows the dimension and the levels coded. The second column shows the percentage (or mean for continuous variables), and the third column shows the standard deviation for the continuous variables. The mean age for respondents is about 51 years (SD=16.1), more than half have a Qualification (53%), but no higher degree, are married (61%) and employed (58.9%). On average, respondents have 0.5 children in the home. The majority of the sample's nationality is British (94%) and completed an interview last wave (81%). As to respondents' cognitive

abilities, the average number of words for the cognitive word test score is 13.1 words (SD=5.1), and more than half correctly completed the subtraction cognition test (61.3%).

The interviewers' average age is 58.6 years (SD=8.2), and similar to respondents, the majority are of British nationality (93.1%). On average, interviewers have six years of experience (SD=3.8) and complete an average of 2.3 interviews per day (SD=0.7). Interviewers also rate the respondent's understanding during the interview as excellent (64.9%) or good (32.7%) and the majority of respondents as having no resistance (86.3%).

Table 2.3 Descriptive Statistics for Respondent, Interviewer, and Interview Context

Respondent Characteristics (n=168)	**%/mean**	**S.D.**
Age	50.8	16.1
Education		
No Qualification (i.e., no high school diploma)	18.5	
Qualification, but Less than Degree	53.0	
Has Higher Degree	28.6	
Married	61.9	
Employed	58.9	
No of Own Children in Home	0.5	1.0
Non-British	6.0	
Cognition Word Score	13.1	5.1
Cognition Subtraction Correct	61.3	
Completed Interview Last Wave	81.0	
Interviewer Characteristics (n=80)		
Age	58.6	8.2
Non-British	8.9	
Experience	6.0	3.8
Average Number of Interviews per Day	2.3	0.7
Interview Context (n=168)		
Interviewer Assessment of R's Understanding		
Excellent	64.9	
Good	32.7	
Fair	2.4	
Interviewer Assessment of Resistance		
No Resistance	86.3	
Soft	8.3	

Moderate	4.2	
Firm	1.2	
Other Present	34.5	
Number of Interviews Same Day	2.3	1.1

Analysis Methods

The first step for the analysis was to assess the relationship between the question characteristics and major deviations. A Rao-Scott Chi-Square test statistic was used to determine significant associations (using SAS 9.4). Consideration was given to including minor deviations in the analysis, but we argue that minor deviations are unavoidable (i.e., to err is human), but do not change the meaning of questions. Conversely, major deviations are more likely to be intentional and do change the meaning of questions. From a data quality perspective, major deviations should be the focus, and thus it is used as the dependent variable in all analysis.

The second step was to run a multi-level model with respondent, interviewer, and interview context level variables to explore the above relationships in more depth. Given the hierarchical nature of the data (question within respondents within interviewers), a multi-level model allows group effects to be accounted for by including variables that measure group characteristics that may influence individual outcomes (i.e., the question characteristics). Multi-level modeling will give correct standard errors and a correct estimate of between-group variance (Steele, 2008).

Model Specification

The variable Changed (i.e., interviewer's question-reading deviation changed the meaning of the question) is the outcome variable, with the question characteristics (including the continuous variables question order (i.e. the order in which the questions were administered to the respondent), word count, question difficulty, and place in series) as predictors and the respondent and interviewer as control variables. The data has a three-level structure where *i* questions (Level 1) are nested within *J* respondents (Level 2) nested within *K* interviewers (Level 3). To account for the clustering effect within respondents and interviewers, a logistic cross-classified multi-level model is used to assess the relationship between the predictor variables and the outcome measure, Changed. The outcome, Changed, is cross-classified at all three-levels.

The model is specified as:

$$logit(\pi_{ijk}) = \beta_{0jk} + \sum_{a=1}^{46} \beta_a x_{ajk}$$

where

$\{x_{aijk}\}$ are question characteristics (a = 1, … , 25), respondent characteristics (a = 26, … , 42) or interviewer characteristics (a = 43, … , 46).

$\beta_{0jk} = \beta_0 + \nu_{0k} + u_{0jk}$
$[\nu_{0k}] \sim N(0, \Omega_\nu) : \Omega_\nu = [\sigma^2_{v0}]$
$[u_{0jk}] \sim N(0, \Omega_u) : \Omega_u = [\sigma^2_{u0}]$

In the model, logit (π_{ijk}) is the predicted log-odds that y = 1 (i.e., the interviewer's question-reading deviation changed the question's meaning). The terms ν_{0k} and u_{0jk} represent the level 3 (interviewer effects) and level 2 (respondent and interview context), respectively. The terms σ^2_{v0} and σ^2_{u0} represent the *unexplained* variance for level 3 and level 2, respectively. The β_a terms represent the additive effect of a 1-unit increase in the dependent variables on the log-odds of the interviewer changing the meaning of the question after adjusting for the group effect of u_{jk} and v_k. However, exponentiating the β_a terms, provides us the odds ratios, interpreted as the

multiplicative effect of a 1-unit increase (or for categorical variables, comparing the measure to the reference category) on the relative odds of an interviewer changing the meaning of the question after adjusting for the group, or clustering effect of the questions within respondents within interviewers. β_a . Continuous variables are centered at the grand mean. The model is estimated using MCMC in MLwiN 3.01.

2.4 Results

Two-way Tables

The results show (see Table 2.4) all question characteristic variables are significantly associated with major question-reading deviations, with one exception, confirming past information. Examining the structure type of question characteristics first, gate questions (17.4%) and gate follow-up (16.3%) are more likely to have interviewers change the meaning of the question than other types of questions (9.1%). For series questions and questions with common stems, interviewers are more likely to change the question's meaning than when these characteristics are *not* present.

Table 2.4. Two-Way Table Question Characteristics by Changed Variable (n=10345)

Question Characteristic	Sample n=10386	% Major Deviation
Structure		
Gate or Independent Question***		
Gate	2613	17.4
Follow up Question	2554	16.3
Other	5178	9.1
Part of Series of Questions***		
Yes	3967	4.8
No	6378	18.1
Stem***		

Yes	1933	5.1
No	8412	14.8
Response Options Read in Text†**		
Yes	3407	11.8
No	5969	13.8
Type of Response†***		
Yes/No	2481	21.6
Select one	4736	8.6
Select all that apply	443	17.6
Scale	670	5.7
Open ended	1046	15.6
Number of Response Options†***		
0 (i.e., open ended)	1046	13.9
2	3156	19.4
3 to 5	2852	8.2
6 to 7	1133	6.4
8+	1189	13.6
Content		
Type of Question***		
Demo/Factual	1825	25.2
Behavioral	4301	12.5
Attitudinal/Belief	3250	5.6
Intro/Instruction	969	12.4
Confirming Past Wave Information		
Yes	1086	13.1
No	9259	13.0
Double Barreled**		
Yes	956	10.4
No	9389	13.2
Time Reference***		
Yes	939	16.5
No	9406	12.6
Sensitive Question***		
Yes	1951	17.7
No	8394	11.9
Question Aids		
Optional Text***		
Yes	1152	7.4
No	9193	13.7
Definition or Example Given***		
Yes	1329	31.8

No	9016	10.2
Showcard†***		
Yes	2380	5.5
No	6996	15.6
Question Help***		
Yes	1132	27.5
No	8244	11.1

Rao-Scott Chi-Square Test
*p<0.05
**p<0.01
***p<0.001
†Conditional on having a response option

Looking at the response option structures, questions where the response options are not read in the question text have a slightly higher percentage of interviewers changing the meaning of the question than when the response options are read in the question, 13.8% compared to 11.8%. One possible explanation for this result is that interviewers may view questions that have the response options in the text as essential pieces of information for the respondent to answer the question.

For the type of response options, Yes/No response options have the highest percentage of interviewers changing the meaning of the question, 21.6%, followed by Select All the Apply (17.6%), Open-ended (15.6%), Select One (8.6%) and Scale having the lowest percentage (5.7%). As for the number of response options, the results do not follow a linear pattern. Questions that have two response options have the highest percentage of interviewers changing the meaning of the question, 19.4%, but the next highest category is questions that are open-ended or have zero response options (13.9%), followed by 8+ response options (13.6%), then 3 to 5 response options (8.2%), with 6 to 7 response options having the lowest percentage, 6.4%. The results align with the type of response option results; Yes/No (i.e., two response options) and opened questions have a higher percentage of change. However, the other number of response

options categories (i.e., 8+ response options, 3 to 5 response options, 6 to 7 response options) suggests that when the number of response options number is 3 to 7 options, interviewers are reading the questions with no or minor deviations. Major deviations occur when the number of response options exceeds seven. If interviewers want to speed up the interview, they may not be reading all the response options. However, in chapter 1, the data shows that interviewers are reading the response options. The interviewer is changing the wording at the beginning of the question. One explanation is that interviewers may see it more important to read all the response options than to read the questions exactly as worded, or that the respondent is interrupting the interviewer because they have a threshold as to how long they will wait for response options to be read.

Turning to the content question characteristics, the type of question, demographic or factual questions have the highest percentage of question meaning change, 25.2%, followed by behavioral (12.5%), introduction or instructions (12.4%) with attitudinal questions with the lowest percentage of change, 5.6%. The attitudinal questions are less likely to have definitions or examples and fewer words than the other question types; thus, interviewers may not feel the need to shorten the questions.

Double-barreled questions have a lower percentage of change than questions that are not double-barreled, 10.4% compared to 13.2%. Interviewers may feel like they cannot change the question if the double-barreled items are distinct.

For time reference, interviewers engage in major deviations at a slightly higher rate than when the questions have a time reference than questions that do not have a time reference, 16.5%

compared to 12.6%. This is most likely because interviewers are omitting time references about half the time (49.8%) when the questions have a time reference. For sensitive questions, the findings are as expected; sensitive questions have a higher percentage of change than non-sensitive questions, 17.7% compared to 11.9%. Interviewers may be changing the wording of the question because they are uncomfortable asking sensitive questions.

Examining the questions that have aids, the results show that some types of aids may hinder the interviewer from reading the question verbatim more than other types of aids. When a question has optional text, interviewers make fewer deviations that result in change than when the question does not have optional text 7.4% compared to 12.7%. This could be because the interviewers are not required to read the optional text, and they perceive it as already shortening the question when they do not read the optional text. However, when a question has a definition or an example that the interviewer is required to read, interviewers are changing the meaning of the question at three times the rate than when the question does not have a definition or example, 31.8% compared to 10.2%. As earlier hypothesized, interviewers may feel like definitions or examples are optional, given that some questions make them optional (by putting the text in parenthesis).

Looking at showcard and question help text, providing a showcard resulted in fewer question-meaning changes than when there is no showcard, 5.5% compared to 15.6%), which suggests showcards not only aid the respondent but also aid in interviewer in reading the question verbatim. However, providing question help text has the opposite effect. When question help text is available to the interviewer, the interviewer engages in more deviations that result in question

meaning change than questions that do not have this feature, 27.5% compared to 11.1%. One possible explanation for this result is that interviewers may feel like they do not have to read the question verbatim because if they run into trouble (e.g., the respondent does not understand the question), they can offer the help text. Another hypothesis may be that the interviewer is trying to incorporate the help text into the question but inadvertently changes the question's meaning.

Multi-level Model Results

The model results show (see Table 2.5) that after controlling for respondent and interviewer characteristics, many question characteristics retain their significant association with question meaning change. Also, many of the question characteristics retain their significant association even when word count and difficulty of question are accounted for. Word count and difficulty of question are significantly associated with question meaning change, while place in series and the number of response options are not.

There is a significant intra-class correlation at both the respondent and interviewer level. Since the level 1 variance is fixed and non-constant for a logit multi-level model, the level 2 and level 3 intra-class correlation (ICC) can be approximated if the level 1 variance is set to a standard logistic distribution, 3.29 (Jones and Subramanian, 2017). The ICC for level 2 (i.e., respondents) is 0.186, and level 3 (i.e., interviewers) is 0.233, indicating that 18.6 percent of the variance is due to the respondent and 23.3 percent is due to the interviewer.

Table 2.5. Model Coefficients, S.E. and Odds Ratios Predicting Question-Reading Deviation

	Est.	S.E.	exp (β)
Fixed Effects			
Constant	-4.878	1.068	
Question Characteristics			
Structure			
Order in Questionnaire	0.003	0.000	1.003*
Gate Questions (ref=Independent Question)			
Gate	0.293	0.123	1.340*
Follow-up	0.521	0.130	1.684*
Part of Series	-0.305	0.229	0.737
Place in Series	-0.105	0.043	0.900*
Double-barreled	0.061	0.161	1.063
Common Stem	-0.332	0.208	0.717
Response Options Read in Question	1.419	0.193	4.133*
Type of Response (ref=Other)			
Y/N	0.076	0.151	1.079
Select 1	-0.941	0.225	0.390*
Select all	-1.120	0.324	0.326*
Scale	-0.850	0.444	0.427*
Content			
Word Count	0.006	0.003	1.006*
Difficulty (FKG)	0.058	0.010	1.060*
Type of Question (ref=Intro/Instruct)			
Attitude	-0.272	0.253	0.762
Behavioral	0.534	0.219	1.706*
Demo/Factual	0.884	0.239	2.421*
Number of Response Options	-0.036	0.023	0.965
Confirming Past Wave Information	-0.007	0.149	0.993
Sensitive Question	-0.093	0.117	0.911
Question Aids			
Optional Text	-1.619	0.177	0.198*
Definition/Example	1.857	0.128	6.404*
Time Reference	0.443	0.142	1.557*
Showcard	0.470	0.216	1.600*
Question Help	0.405	0.119	1.499*
Respondent Characteristics			
Age	0.002	0.011	1.002
Education (ref=Has Higher Degree)			
No Qualification	0.446	0.429	1.562
Less than Degree	-0.314	0.310	0.731
Married	-0.670	0.296	0.512*
Employed	0.425	0.288	1.530*

No of Children in Home	-0.054	0.150	0.947
Non-British	0.573	0.459	1.774
Cognition Word Score	-0.015	0.015	0.985
Cognition Subtraction Correct	-0.281	0.246	0.755
Completed Interview Last Wave	-0.381	0.309	0.683
Interviewer Characteristics			
Age	-0.004	0.020	0.996
Non-British	1.049	0.457	2.855*
Experience	-0.035	0.036	0.966
Average Number of Interviews per Day	0.241	0.264	1.273*
Interview Context			
Interviewer Assessment of R's Understanding (ref=Excellent)			
Good	-0.129	0.303	0.879
Fair	-0.064	0.883	0.938
Interviewer Assessment of Resistance (ref=No Resistance)			
Soft	0.071	0.479	1.074
Moderate	0.635	0.680	1.887
Firm	-0.540	1.138	0.583
Other Present	-0.132	0.284	0.876
Number of Interviews Same Day	0.009	0.110	1.009

Random Effects	**Var (Constant)**	**ICC**
Level: Interviewer	1.321 (0.418)	0.233
Level: Respondent/Interview Context	1.052 (0.306)	0.186

*p<0.05

Results for Question Structure

Examining Question Structure variables first, whether or not response options are read in the text has the highest odds (4.133) for increasing major question-reading deviations than response options not read in the text, holding all other variables constant, including question length. One possible explanation is that for questions that have the response options read at the end of the question, respondents may be cutting off interviewers when they hear a response that fits their

response, and interviewers stop reading the rest of the question, leaving out important information or better fitting response options.

As expected, gate questions (1.340) and gate follow-up questions (1.684) also increase the odds of major deviations than independent questions (i.e., neither gate nor follow up questions). Interviewers may be intentionally changing the wording to induce an answer that avoids the follow-up questions. As for the gate follow-up questions, interviewers may not be as familiar with these questions, as they are not asked of everyone, and making errors in reading the question. However, interviewers recognize that follow-up questions lengthen the interview and intentionally take shortcuts to speed up the interview. Another explanation is that the respondent has answered the follow-up question when answering the gate question, so the interviewer stops or skips reading the follow-up question.

The order in which the question is administered to the respondent has an increase in odds of a major deviation (1.003). Interviewers may feel the pressure (whether from themselves or external forces) to speed up the interview as they progress. However, the later deviations could simply be from interviewer fatigue, and people make more errors when they are tired. Whether or not the question is part of a series does not impact major question-reading deviations, but later questions in a series decrease odds (0.900). This finding conflicts with Presser and Zhao's (1992) finding that being part of a series increases deviations. However, they did not differentiate between minor and major deviations, and this study focuses only on major deviations and has a limited set of control variables that could account for the difference.

All but one of the types of response options (i.e., Yes/No was not significant) also decreases the interviewer's odds of making major deviations relative to 'Other' (i.e., open-ended or no response options) type of response options. This result supports previous findings that open-ended questions increase question-reading deviations (Bradburn et al., 1979; Cannell & Robison, 1971; Mathiowetz & Cannell, 1980; Oksenberg, 1981).

Along with questions that are part of a series, double-barreled and questions with common stems are no longer significant after accounting for the data structure and controlling for other question, respondent, and interviewer characteristics.

Results for Question Content

Word count increases the interviewer's odds of changing the meaning of the question, 1.006, which means there is a 6% increase in the odds of the interviewer will make a major deviation for 10 additional words in the question text. Also, as the question's difficulty increases so does the odds of a major deviation, 1.060 which means there is a 60% increase in the odds the interviewer will make a major deviation for every 10 points the question’s difficulty increases. The finding for word count and question difficulty support previous findings and follow reason. If interviewers are looking to speed up the interview, one sure way is to make questions shorter or skip questions. Also, as questions increase in difficulty, interviewers may feel the need to 'help' respondents more or simply find the question more challenging to read verbatim.

For the type of question, when compared to an introduction or instructional questions, demographics questions increase the odds of question meaning change by 2.421 and behavioral

questions by 1.706. Attitudinal questions are no longer significant after accounting for the data structure and controlling for other question, respondent, and interviewer characteristics. After examining the results, one possible explanation for demographic and behavioral questions increasing the odds of major deviations is that both types ask for factual or quasi-factual types of information. Interviewers, especially interviewers conducting longitudinal interviews, may feel they 'know' the respondent from preloaded data or case notes and feel lesser of a need to read the question verbatim than in cross-sectional or one-off surveys. However, confirming past wave information does not significantly increase the odds of major question-reading deviations. Therefore, there may be other explanations as to why interviewers are more likely to make more major deviations when administering demographic and behavioral questions.

Along with attitudinal questions, sensitive questions are no longer significant after accounting for the data structure and controlling for other question, respondent, and interviewer characteristics. The number of response options is not significant in the model.

Results for Question Aids

All but one of the question aids increase interviewers' odds of making major deviations; optional text decreases the odds. Questions with definitions or examples have the highest odds for increasing question meaning changes (6.404) out of all question characteristics. As stated previously, word count was controlled for in the model, so it is not a matter of longer questions. One explanation could be that definitions and examples often appear as optional text. When interviewers see definitions or examples that are intended to be read, interviewers may incorrectly infer that definitions and examples do not matter and administer the question without

reading them. These results show questionnaire designers should use caution when using definition and examples; making some required and other not, may send the interviewer mixed signals on the importance of reading them.

Questions that have time references, use showcards, and offer question help text increases the odds of interviewers making question-reading deviations, by 1.557, 1.600 and 1.499, respectively. Along with the definition/example results, these results suggest that these question aids may be doing more harm than good, at least in terms of whether or not the interviewer reads the question verbatim. Interviewers may be trying to improve 'difficult' questions; questions with interviewer aids are more likely to have some anticipated difficulty. However, question difficulty was controlled for, so it may be some other reason. One possible explanation is that interviewers may feel lesser of a need to read the questions verbatim because they have a 'backup' if the respondent has difficulty answering the amended question.

As mentioned previously, optional text decreases the odds (0.198) of interviewers making major deviations than questions that do not have optional text. One possible explanation is interviewers perceive omitting optional text as already shortening the question, thus reading the question without any major deviations. Conversely, questions without optional text increase the odds of deviations by about six times. Another possible explanation is that interviewers may misunderstand why questionnaire designers use optional text for some questions but not others. Perhaps interviewers believe that even though a question does not have an indicator for optional text (e.g., parentheses), omitting text from said questions is allowable since other questions allow

text to be omitted. Trainers may need to expand on why some questions have optional text while others do not, emphasizing the importance of reading questions with no optional text.

Results for Respondent and Interviewer Characteristics

Interestingly, only two of the respondent characteristics is significantly associated with question meaning change; married respondents have a decrease in the odds that interviewers deviate relative to those who are not married (0.512), while being employed increases the odds, 1.530. However, the other 'busyness' indicator, number of children in the house show no effect. Perhaps having two adults in the household allows married people to allocate sufficient time for the interview process (e.g., one parent takes care of the children while the other completes the interview), but employed people may have a shorter time frame available to complete the interview than non-married people.

Similarly, only two of the interviewer characteristics is significantly associated with question meaning change; interviewers who are non-British increase the odds (2.855) that the interviewer will make a major deviation. This result could be due to a language familiarity issue. However, native language and language skill level can vary for both non-British (e.g., non-British but originate from an English speaking country) and British (e.g., second generation immigrants who speak a non-English language in the home), it is difficult to link nationality to a language. One should include the interviewer's native language or language ability to investigate this further if the variable is available. The other interviewer characteristics that increases the odds of major deviations is the average number of interviews per day the interviewer completes (1.273). This

result suggests that this field metric should be monitored. Interviewers who have a higher average than the interviewer pool should be flagged and their interviews reviewed by quality control staff.

While this study's focus is question characteristics, the rather sizeable ICCs for level 2 and level 3 suggest there are unaccounted respondent and interviewer characteristics that may contribute to question-reading deviations. Perhaps respondent characteristics that are more thought of as behavioral, personality traits, and other cognitive abilities that were not measured in the IP survey would be better suited to this analysis. For instance, a measure of agreeableness may make a better prediction of a respondent willing to sit patiently through an interview than whether or not the respondent is employed. Similarly, interviewer characteristics that are more about the interviewer's behavior and personality may provide further insight into why interviewers engage in question-reading deviations. Future research in this area should incorporate some of these traits and abilities.

2.5 Conclusions

This study found interviewers engaged in major question-reading deviations in 13% of the questions asked. Using a multi-level model, this study found that of the 19 question characteristics examined, 16 are significantly associated with major question-reading deviations, even when controlling for respondent and interviewer characteristics. Overall the results suggest question structure and question aids tend to have higher odds of the interviewer making a major deviation than question content. The characteristics that have the highest odds of interviewer question-reading deviations are questions that have definitions or examples (6.404), questions that have response options read in the question text (4.133), and demographic questions (2.421).

Results suggest that some question aids (definitions, showcard, and question help) may be doing more harm than good; if interviewers are changing the meaning of the question, then using a question aid to help respondents becomes moot. Interestingly, only two of the respondent characteristics, marital status (respondents who are married have a decrease in odds relative to those who are not married) and employment status (respondents who are employed have an increase in odds relative to those who are not employed), Similarly, only two of the interviewer characteristics, nationality (interviewers who are non-British increase the odds by 2.8) and the interviewer's average number of interviews per day (as the average number of interviews increases, the odds of a major deviation increases by 1.3), were significant.

Although there is more research to be done in this area, there are practical implications one can take away from this study. First, questionnaire designers should try to limit the characteristics shown to increase major question-reading deviations. In particular, questionnaire designers should take into account that questions with definitions or examples are less likely to be read verbatim than questions that do not have this feature. Likewise, questionnaire designers should be aware that interviewer deviations are more likely for questions where the response options are read as part of the question. While it may not be feasible to avoid these features or other characteristics that have been shown to increase major deviations entirely, questionnaire designers could use other techniques to reduce the effect. One such technique could be to insert on-screen reminders that reading the question verbatim is essential and required. Similarly, when training interviewers, instructors may want to draw attention to questions that have shown to increase question-reading deviations and convey the importance of reading the questions

verbatim. Trainers could illustrate this by giving examples of how even the slightest word changes can impact data quality.

Limitations

This study does fill a gap in the literature for this topic; however, there are limitations. First, this study uses observational data, not experimental data, the study cannot fully control for all the characteristics of questions. For example, certain characteristics may not exist in the data, and some combinations may be confounded. Other question, respondent, and interviewer characteristics may also play a role in question-reading deviations, and thus, change the results of this study may change. Second, the behavior coding was only performed on a subset of the interview recordings. Due to technical and administrative difficulties, recordings were not available for all of the interviews. The interviews that were not recorded may be qualitatively different from those recorded. One could argue that interviewers who engage in more question-reading deviations may not want to be recorded.

Question-Reading Deviations and Data Quality

Abstract

In standardized interviewer-administered surveys, the interviewer is tasked with reading all questions exactly as worded. However, research has shown interviewers go off script, engaging in both minor and major deviations. Researchers argue that major deviations, those that change the meaning of the question, increases measurement error. However, there have been very few studies that evaluate whether or not this assumption is accurate. Those studies that have assessed interviewer question-reading deviations have reported mixed findings. Results from these studies show deviations, in some cases, do increase measurement error, while other studies have shown question-reading deviations have no impact on measurement error, or in some cases, actually decrease measurement error. The data from these studies come from either lab settings or CATI surveys, where research has shown the rate and type of deviations are much lower than fielded face-to-face interviews. Hence, there is still much debate on how or if interviewer question-reading deviations affect measurement error. Further, it is unknown how question-reading deviations affect measurement error in face-to-face surveys.

To evaluate question-reading deviations and data quality in face-to-face surveys, this study used interview recordings, paradata, and survey data from Wave 3 of the Understanding Society Innovation Panel (IP). Interviews were behavior coded on whether the interviewer read questions as verbatim or committed a minor deviation or major deviation. Several measures are used to assess data quality, including item non-response and differences in distributions for questions that are read verbatim (or have minor deviations) and questions that have major deviations. In addition, this study exploits several IP Wave 3 experiments on question formation (e.g., branching and presence of showcards) to evaluate whether or not the measurement error (i.e., differential response distributions) found for different question formations can be partially attributed to interviewer question-reading deviations.

3.1 Introduction

In standardized interviewer-administered surveys, the interviewer is tasked with reading all questions exactly as worded. However, research has shown interviewers go off-script, engaging in both minor and major deviations (Ackermann-Piek and Massing, 2014; Belli and Lepkowski, 1996; Cannell, Lawson, and Huasser, 1975; Haan, Ongena and Huiskes, 2013; Mathiowetz and Cannell, 1980; Oksenberg, Cannell and Kalton, 1991; Schumann and Presser, 1997). Researchers argue that major deviations most likely change the meaning of the question, thus increasing measurement error (Groves et al., 2009; Krosnick, Malhotra, and Mittal, 2014; Rugg, 1941; Schuman and Presser, 1996). However, there have been very few studies that evaluate whether or not this assumption is accurate. Those studies that have assessed interviewer question-reading deviations have reported mixed findings, with some studies finding negative associations with data quality (Schumann and Presser, 1997), others finding a positive association (Dykema, Lepkowski, and Blixt, 1997; Haan, Ongena, and Huiskes, 2013) and still others find both positive and negative associations (Belli et al., 2004).

The data from two of the studies examining errors relating to interviewer question-reading deviations (Schumann and Presser, 1997; Haan, Ongena, and Huiskes, 2013) use data from CATI surveys, where interviewer behavior can be quite different when a supervisor or co-worker is in close proximity. Research has shown the rate and type of deviations are much lower than fielded, face-to-face interviews; telephone interviews range from a low of 4.6% (Mathiowetz and Cannell, 1980) to a high of 36% (Cannell, Lawson, and Huasser, 1975), and in face-to-face interviews, these can be as high as 84% (Ackermann-Piek and Massing, 2014). The other study

assessing these deviations (Dykema, Lepkowski, and Blixt, 1997) used data from a face-to-face validation survey, and limited analysis to 10 questions, which the authors acknowledged limitation and further state research is needed. Hence, there is still much debate on how or if interviewer question-reading deviations affect measurement error, especially in face-to-face interviews.

To evaluate question-reading deviations and data quality in face-to-face surveys, this study used interview recordings, paradata, and survey data from Wave 3 of the Understanding Society Innovation Panel (IP). Interviews were behavior coded on whether the interviewer read questions as verbatim or committed a minor deviation or major deviation. To assess data quality, several measures are used, including item nonresponse, question timing, and exploits several IP Wave 3 experiments on question structure (e.g., branching and presence of showcards) to evaluate whether or not the measurement error (i.e., differential response distributions) found for different question structures can be partially attributed to interviewer question-reading deviations.

3.2 Background

Interviewers can and do affect data quality, particularly measurement error (West and Blom, 2017). To minimize this measurement error, organizations train interviewers in standardized interviewing techniques, which is the most widely used interviewing style (Groves et al., 2011). In standardized interviewing, interviewers are instructed to strictly follow all study protocols so each respondent receives the same 'treatment', thus reducing the variability that can arise from having different interviewers interviewing the study's target population. The core and widely

supported principle of standardized interviewing is reading all questions verbatim (Groves et al., 2011).

Reading questions verbatim is widely supported because question-wording experiments have shown that even slight wording changes can change the meaning of the question (Groves et al., 2011; Krosnick, Malhotra, and Mittal, 2014; Rasinski, 1989; Schumann and Presser, 1996). If meaning is changed, as these authors suggest, then deviations change the stimulus respondents are reacting to, and there is no guarantee that responses are comparable. However, these studies are question-wording experiments where the researcher manipulates the questions to test different versions of the question. Interviewers who deviate in the field are not given an example or a scripted alternate version on how to change the wording – something else prompts them to change the question's wording.

What that "something else" is, is not understood very well, but regardless, the fact remains that interviewers do make question-reading deviations. In some cases, interviewers are simply making reading errors. In other cases, researchers hypothesize question-reading deviations run from trying to help respondent comprehension (Schober and Conrad, 2002), to signaling the respondent they are listening (Haan, Ongena and Huiskes, 2013; Ongena and Dijkstra, 2006), to intentionally falsifying data for their gain (Winker, 2016). Interviewers may try to help respondents because they perceive a question as too complex, or past experience administering the question precisely as worded led to respondent comprehension problems, so they are trying to 'help' the respondent (or the next respondent) better understand the question.

For example, in an interview, suppose the interviewer-administered read the question exactly as worded, but the respondent had comprehension issues and asked the interview for clarification. After giving the respondent clarification, the respondent gave a codable answer. Whether or not the clarification did indeed improve the quality of the answer is not clear, but the interviewer was able to meet their objective: obtain a codable answer. In the next interview, the interviewer remembers the comprehension issue with the previous respondent, so instead of reading verbatim, the interview works in the clarification into the initial reading of the question text. One could argue an interviewer, especially a well-trained and experienced interviewer, has developed the skills to recognize cognitively challenging questions, and their adaptation to the question (e.g., omitting, adding, or substituting words) is improving data quality.

Question-reading deviations may be driven by another source – the question's characteristics. The literature here is also sparse, but several studies found deviations can increase for open-ended question vs. closed-ended questions (Bradburn, Sudman, Blair, Locander, Miles, Singer and Stocking, 1979; Cannell and Robison, 1971; Mathiowetz and Cannell, 1980), longer questions vs. shorter questions (Bradburn et al., 1979; Presser and Zhao, 1992), and for questions that are part of a series (Presser and Zhao, 1992). The above studies were limited in terms of how many question characteristics they tested. However, in Chapter 2 of this dissertation, 19 question characteristics were examined, and 16 of these characteristics were significantly associated with major question-reading deviations, even when controlling for respondent and interviewer characteristics. Of the 16 characteristics, interview aids (e.g., showcards, definitions, and help text) were shown to have the highest impact on increasing the odds of interviewer deviations.

However, the study did not examine how these deviations and the interaction between deviations and question characteristics might affect data quality.

While the cause of these deviations deserves more study, there is also a lack of understanding as to the impact these deviations have on data quality, particularly in general, 'natural' (i.e., not experiments pre-testing questions) survey environments. There are a few studies that examine deviations and data quality, and the findings of these are mixed. Haan, Ongena, and Huiskes (2013) report deviations decrease measurement error, hypothesizing that deviations are not always a negative interviewer behavior. Some changes in question reading may increase both the cohesion and the coherence within the interview, thus having a positive effect on data quality. Schumann and Presser (1997) report the opposite - deviations can increase measurement error when evaluating five question-wording experiments. However, the authors acknowledge that the experiments were designed in such a way (i.e., manipulating the wording with terms that should induce differences in responses) that they expected a wording effect. In a validation study, Dykema, Lepkowski, and Blixt (1997) show that question-reading deviations have no "consistent" impact on measurement error.

A proposed alternative to standardized interviewing is conversational interviewing, which gives the interviewer the freedom to formulate questions in their own words, tailoring the interview to the respondent in order to achieve the goals of the interview. While there is some debate about how much freedom interviewers should have in conversational interviewing, a set of research that allows conversational interviewing only after initially reading questions verbatim has shown improvement in data quality (Schober and Conrad, 1997; West, Conrad, Krueter and Mittereder,

2018). However, the 'conversational' technique comes into play in how the interviewer follows up the question (e.g., probing and clarification) when the respondent fails to give a codable answer to the initial reading of the question. Given this interviewing style instructs the interviewer to read questions verbatim initially, one could argue these researchers believe there may be a risk to data quality when interviewers go off script in the initial reading of the question. Although these studies show that conversational interviewing can improve data quality, the interviewers were trained in conversational interviewing in both of these studies. One could argue that in studies where the interviewers are not trained in conversational interviewing, the interviewers may not have the knowledge on how to go off-script in a way that they do not bias the respondents' answers.

Additional research has indeed suggested that a more conversational form of interviewing improves data quality. In particular, several studies show that standardized interviews produce lower quality data than conservational interviews using an event history calendar (EHC) in the Panel Survey of Income Dynamics (PSID) (Belli et al. 2004; Belli, Bilgen and Al Baghal 2013; Belli et al. 2016). This improved data quality may occur due to conversational techniques being more natural forms of communication (Houtkoop-Steenstra 2000) and less likely to flout maxims of conversation, which are important in understanding survey outcomes (Schwarz 1996). More importantly, the use of conversational techniques in the EHC are more aligned with the varied structures of autobiographical memory, whereas standardized interviewing relies more on one aspect of such memory (Belli 1998; Belli and Al Baghal 2016). However, the findings for improved data quality come largely from EHC data, where interviewers were trained in

conversational techniques. These results may not hold in a traditional standardized survey, where interviewers deviate in contrast to what their training provides.

Some researchers argue for more flexibility within conversational interviewing for traditional interviews. Haan, Ongena, and Huiskes (2013) argue that giving freedom in the initial asking of the question provided a cohesive interviewing experience to the respondent, which produced higher quality data. However, the authors state many of the question-wording changes were made to "specific interactional functions", meaning the interviewers were not changing the core of the question. One could argue these types of changes (i.e., changes to facilitate cohesion in the interview) may have a positive effect on data quality because the wording changes did not majorly change the core of the question. The study's (Haan, Ongena, and Huiskes, 2013) findings give further evidence that deviations in the initial wording may be acceptable, even advisable if the deviation is made in an attempt to increase data quality. However, more research is need on the type of deviations and its effect on data quality, especially for fielded, face-to-face surveys before interviewers can be trained to know which types of deviations can increase data quality and which types of deviations decrease data quality.

This study attempts to fill that gap in the literature for deviations and data quality by examining the following research questions:

- In face-to-face interviews, do major deviations to question wording reduce data quality?
- In face-to-face interviews, do major deviations to question wording interact with features of the questionnaire to impact data quality?

3.3 Methods

Sample

This study combines paradata derived from audio interview recordings (i.e., interview behavior coded data), the questionnaire (i.e., question characteristic coded data), and survey data from Wave 3 of the Understanding Society Innovation Panel. Understanding Society is a household panel study interviewing 40,000 households in the UK on various social and economic topics. The Innovation Panel (IP) is a separate panel for methodological research (i.e., experiments and testing questions, procedures, and methods in a context similar to the main study) with the results taken into consideration in the development of the next wave's main stage instruments (University of Essex, 2019). The IP uses a multi-stage probability sample with an initial household CAPI interview to determine eligibility and collect household-level information. The target sample size for Wave 1 was 1500 households, and addresses were randomly selected from the Postcode Address File (PAF). Interviews were conducted at 1489 households (59.0% response rate), and 2393 individual interviews were completed, with an 88.9% conditional individual response rate. Respondents who completed an interview at Wave 1 were invited to participate in subsequent waves. For Wave 3, 1525 eligible households were identified, and 1027 household interviews were completed with a response rate of 73.9%. All eligible adults (age 16+) in the household were then selected to complete an individual, face-to-face, computer-assisted personal interview (CAPI). Conditional on the household response rate, the individual response rate was 82.2%, for a total of 1621 completed interviews. The average interview length was 37.5 minutes, and interviewers are instructed to read all questions verbatim. Selected sections of the interview were recorded with the respondent's permission (72% consent rate).

However, due to procedural and technical difficulties, only 820 interview recordings were available for analysis.

Behavior Coding Sample

This study selected a subset of the available interview recordings (n=820). The recordings were stratified by interviewer (n=80), and interviews were randomly selected from each interviewer. In all, 314 recorded interviews were selected for behavior coding. Questions from the Wave 3 IP Questionnaire were included in the dataset on the following criteria: 1) if the question was intended for the interviewer to read the question aloud to the respondent; 2) did not have varying number of words based on the previous answer or the respondent characteristics (i.e., fills), 3) were administered to both males and females (e.g., omitted fertility questions); 4) had a one-to-one match with timing file questions (i.e., did not loop); 5) had the same response options for all regions (i.e., did not include questions that have regional based response options); and 6) the question was recorded. The questions selected for analysis (n=361) were coded for each of the recordings sampled. Because routing through the questionnaire is dependent on respondents' answers, not all questions are asked of respondents. In total, 13,114 questions administrations were behavior coded for analysis.

Behavior Coding Method

The behavior coding was done directly from the audio files (no transcription) by two coders, with a subset of questions double coded. The coding builds on Cannell, Lawson, and Hausser's (1975) behavior coding scheme. The interviewer's first reading of each question was coded as a) question read verbatim, b) contains only minor deviations, or c) contains at least one major

deviation. Building on Cannell et al. (1975), explicit rules were created to evaluate if the deviation was minor or major (see Chapter 1), with the primary distinction being the assumption that minor deviations most likely do not change the meaning of the question but major deviations are likely to change the meaning of the question. We are interested in deviations that are thought to change the meaning of the question only, so we collapsed the variable into a binary variable Change of No Change (of the meaning). Coding results show that 11.7% of questions had major deviations (i.e., Change).

The coding frame and plan was developed by the first author, and a second coder was hired and trained on coding and use of this frame. To ensure that coding was consistent across the two coders used in this study, 290 questions were independently coded. The concordance on the change code (0= No change, 1= Minor, 2 = Major) was very high; the kappa statistic was used to test concordance account for chance, and the k =0.93, which is considered "strong" interrater reliability (McHugh 2012). As such, codes are treated in a unified way going forward.

Initially, all questions were coded for 168 respondents (see Chapters 1 and 2). Doing so allowed for greater breadth of data available, particularly for the understanding scope of deviations and timings (Chapter 1) and the variety of question types available for analysis (Chapter 2). This also allows for breadth in analysis for outcomes indicated across all measures (see Data Quality Measures, below). However, to add depth (and power) to the data, additional coding focused on the subset of questions used in the branching and showcard experiments. An additional 141 respondents with recordings on these questions were selected, and all questions used in these experiments were also coded.

Question Characteristic Coding Variables

The dimensions coded and the operationalization of question coding varies from study to study. For example, Mangione, Fowler, and Louis (1992) code questions on four dimensions: Sensitive/Not sensitive; Difficult/Not Difficult; Opinion/Factual; Open/Closed. Presser and Zhao (1992) code questions on four dimensions: Length, Position, Familiarity, and Series. An example of differing operationalization of question coding within dimensions, Mangione, Fowler, and Louis (1992), categorize questions as open or closed questions, while Olson and Smyth (2015) code questions as open-ended text, open-ended numerical, closed nominal, closed-ordinal, yes/no.

The question characteristics were expanded to examine specific question design components (see Chapter 2 for coding methodology). Table 3.1 shows the distribution of the question characteristics. The first column shows the dimension and the levels coded. The second column shows the number of each question character for questions in the Wave 3 IP Questionnaire. The third column shows the percentage (or mean for continuous variables) coded for each question characteristic used in the analysis, and the fourth column is the standard deviation for the continuous variables.

Table 3.1 Descriptive Statistics for Question Characteristics

Question Characteristics	**Selected Questions**	**%/mean**	**SD**
Gate or Independent Question			
Gate	71	25.3	
Follow up Question	164	24.7	

Independent	126	74.7	
Word Count	290	25.8	19.7
FKG Score (Difficulty)	290	8.0	4.3
Type of Question			
Demo/Factual	73	17.6	
Behavioral	151	41.6	
Attitudinal/Belief	124	31.4	
Intro/Instruction	13	9.4	
Double Barreled	24	9.2	
Confirm Past	25	10.5	
Sensitive Question	60	18.9	
Showcard	96	23.0	

Data Quality Measures

We use four data quality indicators, two generally used in other studies, and two that leverage experiments in the IP Wave 3. We select indicators based on possible relevance to the subset of questions audio-recorded and behavior coded. For the more general indicators of data quality ('Don't Know' responses and time), all questions can be analyzed. When analyzing experimental data, we include only the further subset of questions available from this experiment on a similar measurement scale. We do not include introductory text in our analyses; although these are coded as having changed or not, these have no outcome to indicate data quality. All analyses take account of the clustered nature of the data.

'Don't Know' response. 'Don't Know' responses are frequently used as a data quality indicator, as these are treated as item nonresponse (e.g., Krosnick 1991; Al Baghal and Lynn 2015; Wenz, 2021). For the initial comparison of differences, the proportion of 'Don't Know' responses for questions where the wording was changed is compared to those where no change occurred. For multivariate analyses, we use a dichotomous measure for each question, indicating whether a

'Don't Know' response or some other response has been selected. We analyze all questions coded with an outcome indicated. Most of the questions are factual, with a small number being attitudinal (e.g., neighborhood cohesion).

Question Timing

As with 'Don't Know' responses, we analyze outcomes across all coded questions (except introductory texts). Studies that have relied on similar data have referred to response times; however, we more appropriately refer to it as question times. The time of the question is not only a function of respondents' speed of answering but also influenced by how the interviewer conducts the survey (Couper and Kreuter, 2013). The amount of time a question takes is seen by a number of studies as an outcome related to data quality (Yan and Tourangeau 2008; Lenzer et al. 2010; Couper and Kreuter, 2013; Zhang and Conrad 2014; Conrad et al. 2017). However, what is the "right" time to ensure that data quality is unclear. Some studies have explored response times that are too fast for data quality (Zhang and Conrad 2014; Conrad et al. 2017), while others have pointed out how too slow times are likely related to reduced data quality (Yan and Tourangeau, 2008; Lenzer et al. 2010). We do not take a position here as to what is fast or slow. Rather, in line with Couper and Kreuter (2013), we note the importance of question times on data quality while focusing on what impacts differences in times. Specifically, we are interested in how changes in how questions are asked impact question times. Due to the skewed distribution of times, we take the natural log and use this in all subsequent analyses.

Branching Measurement Experiment

Here we use the same experiment analysed in Gilbert (2015). For attitude questions asking for bipolar directional options (e.g., agree/disagree), it is common to ask both direction and intensity at once (e.g., strongly agree, agree, neither agree nor disagree, disagree, strongly disagree). Conversely, branched options first ask for direction only (e.g., agree, neither/nor, disagree), and then for respondents selecting a direction, asked the strength of that direction (e.g., strongly agree or agree). There is some evidence that asking in branched format may be a preferred way to measure these attitudes (Schaeffer and Presser 2003). Regardless, there are differences in response distributions based on whether an unbranched or branched scale is used (Kronick and Berent 1993; Gilbert 2015), and these differences suggest differential measurement processes (and error). In particular, Gilbert (2015) found that using a branched design led to more extreme selections (i.e., strongly) of direction than unbranched.

Therefore, we use an indicator of whether the extreme option (strongly agree or strongly disagree) is selected on attitudes of neighborhood cohesion and political efficacy (eight total questions, four for each topic). We explore differences in extreme responses across instances of wording changes or not. In further analyses, in addition to whether the wording was changed, models include both the experimental allocation (i.e., branching or unbranching versions) and the interaction between changes in wording and the experimental allocation. That is, does changing the words in a question affect experimental validity?

Table 3.2 shows the distribution of the branching experiment data, showing that our subsample closely follows the results presented in Gilbert (2015). In particular, we observe more extreme responses for branched questions compared to unbranched questions. This experiment is the one

indicator of data quality that has been used elsewhere from IP3, and we are able to show that our data can reproduce other published work.

Table 3.2. Distribution of Branching Experiment Data

Response	**Branched (n=947)**	**Unbranched (n=903)**
Strongly Agree	10.70	4.22
Agree	11.30	17.57
Neither Agree Nor Disagree	13.84	11.62
Disagree	9.08	12.49
Strongly Disagree	6.27	2.92

Showcard Experiment

We similarly explore the impact of changing the wording on another experiment. In this second instance, we analyze a showcard experiment also conducted in the IP. In face-to-face surveys, showcards are frequently used as an aid, both to communicate response options and to reduce respondents' cognitive burden (Tourangeau et al. 2000). However, if a mode was used lacking a physical presence (i.e., telephone), then data quality comparisons with a face-to-face may be limited due to the impact of differences in available tools, such as showcards. As such, this experiment compared the impact on data outcomes when using showcards or not for a subset of questions in the survey. As with the branching experiment above, differences in distributions would suggest differential measurement and measurement error.

We use the subset of three questions asking respondents how often they talked about political affairs with different groups of people, all asked on a 1-6 scale (1 = Always, 6= Never). We

compare the mean response to these questions first on whether wording was changed or not and then include whether respondents were shown a showcard when asked these questions or not. As with the branching experiment above, we look at whether there is an interaction between changing the wording and experimental allocation (showcard or not).

Analysis Methods

All quality indicators are indicated at the item level, occurring within both respondents and interviewers. As outcomes are nested within both respondents and interviewers, a three-level, cross-classified multilevel model is used for all multivariate analyses (e.g., Yan and Tourangeau 2008). Interviewers are not interpenetrated across the primary sampling unit (PSU), i.e., one interviewer represents one PSU. The inclusion of random effects for the interviewer captures the clustering of PSU. Stratification is not included, but including stratification is expected to reduce variance estimates. Hence, the estimates are likely to be more conservative regarding statistical significance.

Models for 'Don't Know' response and extreme response in the branching experiment are binary outcomes, and logit-link models are used. As such, odds ratios are reported for estimates. Question time (log) and response on the political efficacy scales are modeled as continuous outcomes, and coefficients are presented. For the experimental models, an experimental allocation is included, as well as the interaction between this allocation and changes in wording, to show how changes might impact experiment results. For 'Don't Know' and question time models, examining the measures at the question level allows for including question characteristics in the model described above to further disentangle possible effects for wording changes. However, these question characteristics are not useful for modeling the experiments'

data, as the items are all of the same type and same scale (by design), and so are constant on these indicators. All models include random effects for questions, however, as well as for interviewer and respondent effects.

The respondent and interviewer characteristics included in analyses are the same for all the models. Respondent indicators of sex, unemployment, education, number of children, and being single in the household are included in all models. For educational attainment, those with less than a professional degree are in the baseline educational category, compared to those with a professional or university degree. A proxy measure for the respondent's understanding of the questionnaire comes from the interviewer's subjective rating of the respondent's understanding on a five-point scale. The majority of respondents are rated as having had an "excellent" understanding of the questionnaire. This category is used as the baseline, with comparisons against those having "good" understanding and the combined grouping of "fair"/"very poor" (no respondent in this sample was rated the fourth category, "poor". The latter categories are grouped due to the relatively small proportions given this rating.

The availability of interviewer indicators allows for a possible explanation of interviewer effects beyond what is capture in the random effect variance. The interviewer demographics available from the fieldwork agency include age, sex, and ethnicity. However, a large number of interviewers refused to disclose their ethnicity (21.8%), so interviewer ethnicity will not be considered further. Experience as an interviewer at the fieldwork agency is also included. The average number of daily interviews completed by the interviewer is calculated from the IP data. While daily interviews may indicate effort and success, it may also be an indicator of speeding or fatigue.

There are 13114 individual questions behavior coded overall, which is the initial base for 'Don't Know' and question time analyses. Not every respondent has all experimental questions coded due to access to recordings. Of the total, there are 1850 questions behavior coded and used for the branching experiment analyses, and 960 behavior coded questions used for the showcard experiment analyses. However, missing data occurs for some respondents on one or more predictor variables outlined here. We use list-wise deletion in multivariate analyses, leaving 13003 total questions for 'Don't Know' and question time estimation, and 1819 and 943 questions for branching and showcard experiments, respectively. There are 314 total respondents and 80 total interviewers to be used in analyses. Data are available from 309 respondents in 'Don't Know' and question time analyses, 293 respondents in branching experiment analysis, and 178 respondents for the showcard experiment multivariate analysis. The number of interviewers available for multivariate analyses is not reduced by list-wise deletion but reduced because not all respondents had all questions coded. There are all 80 interviewers for 'Don't Know', question time, and branching experiment analyses, with 76 interviewers in the showcard card experiment.

Respondent and Interviewer Characteristic Variables

Table 3.3 shows the variables described above for respondents and interviews, which are used in the full models to predict data quality. The majority of the sample is female and older, with high percentages of unemployed (unemployed includes retired, students, non-paid care givers, and not seeking work) and those with less than a professional degree. However, nearly 66% had an interviewer-rated "excellent" understanding of the questionnaire, with 29.5% have a "good", with 4.2% having a "fair" or "very poor" understanding. An even higher percentage of

interviewers are women, and the average age is higher than for respondents. Interviews tended to have several years of experience on average, and they completed slightly more than two interviews per day on average.

Table 3.3. Mean/Proportion for Respondent and Interviewer Characteristics

	Proportion/Mean
Respondent Characteristics	
Unemployed	0.424
Single	0.309
Number of Children	0.500
Age	51.01
University Degree	0.161
Professional Degree	0.264
Female	0.576
Good Understanding	0.295
Fair/Very Poor Understanding	0.042
Interview Characteristics	
Interviewer-Age	58.43
Interviewer-Female	0.613
Years as Interviewer	5.99
Average Interviews/Day	2.39

3.4 Results

Descriptive Statistics for Quality Indicators by Changed Status

Table 3.4 shows the descriptive statistics for the quality indicators by changed status. The full sample had very few 'Don't know' responses, and the association between 'Don't Know' and changed is not significant (0.46% vs. 0.69%). However, there is a significant association between question timing and changed, with questions that are unchanged significantly associated with longer timing durations. This finding is expected, as previous research shows that the majority of

major deviations (in this dataset) are due to interviewers omitting words (see Chapter 1). We stated previously that the existing literature is inconsistent on which question timing, shorter or a longer time, is better for data quality. However, when the shorter times are associated with changes in the question meaning, and the majority of the changes can be attributed to omitting question text, then one could argue that shorter question timing has a negative impact on data quality.

Table 3.4. Descriptive Statistics for Quality Indicators by Changed Status

Full Sample	**Changed (n=1537)**	**Unchanged (n=11577)**
% Don't Know	0.46%	0.69%
Mean Question Time (log)	2.11	2.20*
Branching Experiment	**Changed (n=96)**	**Unchanged (n=1754)**
% Extreme	29.17%	23.83%
Showcard Experiment	**Changed (n=43)**	**Unchanged (n=917)**
Mean Scale Response (1-6)	4.35	4.71*

For the branching experiment, there is no significant association between the extreme options and changed. However, for the showcard experiment, we do see a significant association between the mean of the responses and changed with the lower mean for questions that were coded as changed in meaning. The differences in means for changed and unchanged itself does not indicate which is the 'true' mean, but there is a difference, and it is more likely that changing the meaning of the question has a negative impact on data quality; thus, the lower mean may

have more measurement error. Regardless, the differences are suggestive that changing wording changes measurement (for better or worse).

'Don't Know' Response and Question Timing

To more deeply explore the impact of interviewers changing the wording on 'Don't Know' responses and question time alongside question characteristics, multilevel models were run with these as outcomes. Table 3.5 shows the results of these models. Several results are important to note. First, the 'Don't Know' model shows that after controlling for the question, respondent, and interviewer characteristics, the impact of changed wording is not significantly associated with 'Don't Know' responses. Additionally, while this chapter's focus is on changed wording, it is worth noting that eight of the nine question characteristics are significantly associated with Don't Know responses. The most striking association is the odds of a respondent answering 'Don't Know' to a sensitive question is seven times more than a non-sensitive question. Tourangeau et al. (2000) argue that respondents may edit their responses due to embarrassment or hide information from third parties, and this study's finding for Don't Know supports this argument. Also, the odds of a respondent answering 'Don't Know' is almost two and half times likely for an attitude question than a non-attitude question. When respondents answer 'Don't Know' to an attitude question, it may be because they are using "Don't Know" because there is not an explicit "No Opinion" option, or it could be that they do not want to share their opinion with a third party. In either case, "Don't Know" is generally perceived as nonresponse and thus a negative for data quality.

Table 3.5. Models Predicting 'Don't Know' Response (OR) and Question Timing (Log)

	Don't Know (OR)	Time (Log)
Changed	0.654	-0.377*
Question Order	1.001	0.000111*
FKG Score	1.116*	0.0122*
Word Count	0.948*	0.0294*
Showcard	0.502*	0.203*
Attitude	2.289*	0.0317*
Gate Question (Gate)		
Gate Follow-up	1.509	-0.111*
Not Gate	2.241*	-0.0164
Confirm Past	0.153	-0.113*
Double Barreled	0.0502*	-0.0911*
Sensitive	7.202*	-0.0261
R. Age	1.053	0.00347*
Unemployed	1.770	0.0251
Understanding (Excellent)		
Good	1.533	0.00255
Fair/Poor	9.855	0.102
Education (Less than professional)		
University Degree	0.640	0.0407
Professional	0.247	-0.0120
R. Female	1.325	-0.00770
Number Children HH	1.426	0.00421
Single in HH	1.592	0.0548*
I. Avg. Interviews/Day	0.806	-0.0298
I. Female	0.526	-0.0110
I. Age	0.916*	0.00659*
I. Yrs. Experience	1.121	-0.00394
Constant	----	0.983*
Respondent Variance	6.021	0.025
Interviewer Variance	1.760	0.008
n Questions	13003	13003
n Respondents	309	309
n Interviewers	80	80

While changed wording does not have an apparent effect on data quality using Don't Knows, results show that changed wording has a significant negative association with question timing.

Changed question wording leads to shorter question timing (coef. = -0.377, $p < 0.05$) after controlling for the question, respondent, and interviewer characteristics. As stated previously, this finding is not altogether surprising as the majority of deviations are due to the interview omitting words. Shorter question times suggest lower data quality. With the behavior coded data, we can see why there are shorter times; interviewers are omitting words. These changes in wording are to the extent that is thought to change the meaning of the question and is suggestive of the negative effects changes in wording can have on data.

Again, we see that question characteristics also have significant associations with question timing. As question order (coef. = 0.000111, $p < 0.05$), difficulty (FKG Score) (coef. = 0.0122, $p < 0.05$), and word count (coef. = 0.0294, $p < 0.05$) increase, question timing increases. This finding provides further evidence that longer and more difficult questions take longer to administer (Olson and Smyth, 2015). We also found questions with showcards (coef. = 0.203, $p < 0.05$), also increase question timing. This finding is somewhat surprising. Showcards are thought to help the respondent and reduce the time it takes to administer questions (Green, Krosnick, and Holbrook, 2001). However, it could be that the increased time is due to the interviewer adding reminders (that are not scripted) to refer to a showcard or the respondent taking the time to read through the options. So, while the showcard may help the respondent give a codable answer, it comes at the cost of longer question duration timings.

The results also showed longer question timing for attitude questions (coef. = 0.0317, $p < 0.05$). This is the opposite of what Olson and Smyth (2015) found but aligns with previous research (Bassili and Fletcher 1991; Tourangeau et al. 2000; Yan and Tourangeau 2008). Suppose a respondent is being asked about an attitude to a topic that they have never given much thought

(or never thought) to. In that case, the respondent has to recall and retrieve relevant information and integrate it into the topic, which may take more time to formulate an answer than behavioral or demographic questions where the respondent can quickly calculate the answer or 'just knows' the answer. The longer timing durations may reflect better data quality if the longer times are due to the respondent going through the complete response process.

The rest of the question characteristics have a negative association with question timing. Gate follow up questions (coef. = -0.111, $p < 0.05$), confirming past information (coef. = -0.113, $p < 0.05$) and double-barrelled (coef. = -0.0911, $p < 0.05$) have shorter question timings. For the gate follow up questions and the confirming past information type of questions, shorter times may not necessarily mean lower data quality. For example, the respondent is already primed to think about the topic when asked a follow-up question, resulting in less time to retrieve the relevant information to give a codable response. The same could be said for confirming past information; the respondent is not being asked to recall and retrieve anything but instead is presented the relevant information, thus resulting in a quicker response. However, for the double-barrelled questions, the shorter timing may indicate lower data quality, as the respondent may be disregarding one of the references in the question and thus taking shortcuts in the response process.

As for respondent characteristics, the only respondent characteristics that show a significant association is in the question timing model is age and marital status; older respondents (coef. = 0.00347, $p < 0.05$) and single respondents in the household show an increase in question timing (coef. = 0.0548, $p < 0.05$). The finding that older respondents have longer response times supports previous research (Yan and Tourangeau, 2008). However, the latter finding (i.e., single

respondents) is somewhat surprising. One would think the opposite – non-single respondents (i.e., married or partnered) would have longer question timings than single households, as they would possibly have more distractions during the interview. As for data quality, if one prescribes to longer question timing equals better data quality, this data suggests that data from a single household may have better data quality than non-single households.

Branching Measurement and Showcard Experiments

Leveraging the experiments that are unique to the IP, Table 3.6 shows the results for branching and showcard experiment models. There is not a significant impact of changed wording on response outcomes for either experiment. The lack of significance includes both main effects of changed wording and interactions with the experimental allocations. However, the main effect of the experimental allocation is significant in the branching experiment data. Unbranched questions have lower odds of extreme responses than branched questions (OR = 0.198, $p < 0.05$), consistent with other findings (Gilbert 2015). These findings suggest that while questions were coded with a major deviation, the question wording was not changed enough to alter the experiment or that the branching wording had such a strong impact that any deviation was not enough to impact it.

Table 3.6. Models Predicting Extreme Option in Branching Measurement (OR) and Mean Scale Response in Showcard Experiment

	Extreme Option Branching (OR)	Mean Scale Response Showcard
Changed	1.377	-0.171
Unbranched	0.198*	
Changed*Unbranched	1.342	
Showcard		0.0450
Changed*Showcard		-0.327

R. Age	1.031*	0.00292
Unemployed	0.904	0.00763
Understanding (Excellent)		
Good	1.419	0.320*
Fair/Poor	0.755	0.435
Education (Less than professional)		
University Degree	0.979	-0.655*
Professional	0.904	-0.202
R. Female	0.916	0.0277
Number Children HH	0.981	0.106
Single in HH	1.064	-0.109
I. Avg. Interviews/Day	1.128	-0.0708
I. Female	1.010	-0.0689
I. Age	0.997	-0.00698
I. Yrs. Experience	1.003	0.00991
Constant	----	5.123*
Respondent Variance	2.232	0.303
Interviewer Variance	0.041	0.016
n Questions	1819	943
n Respondents	293	178
n Interviewers	80	76

Few of the other variables used are significant predictors of these data quality, either. Age is the only other significant predictor for selecting an extreme response option in the branching experiment. Older respondents have higher odds of selecting an extreme option than younger respondents. This finding supports previous research that older respondents are more likely to shortcut the response process due to declining cognitive abilities and give more extreme responses (both at the low and high end) (Schneider, 2018).

Like the result for the branching experiment, major deviations to the questions was not changed enough to alter the showcard experiment, or the presence of a showcard mitigated the effect of any deviations. Looking at the other variables, those with an interviewer-rated good

understanding of the questionnaire have a higher mean on the political efficacy questions than those with an excellent understanding, and those with a university degree have a lower mean on the political efficacy scales than those with education less than a professional degree. One cannot say whether a higher (or lower mean) mean is an indicator of better (or worse) data quality, but there is a difference in the means.

3.5 Conclusions

Research has shown that interviewers do not always read survey questions as written, which contravenes the desire for standardized administration. However, the impact of these changes in wording on data quality has been researched far less. We add to the research in this area; in this paper, we examine data quality when interviewers engage in major changes in question-wording. We explore data quality through frequently used indicators as well as leveraging the experimental nature of the IP. In particular, we evaluated the impact of changed wording on 'Don't Know' responses, question timing, and response distributions for two experiments (a branching experiment and a showcard experiment). Initial differences in bivariate distributions show that questions with changed wording have faster question times and have a lower mean response on political efficacy scales. These initial findings suggest that interviewer deviations have a negative impact on data quality.

However, after controlling for the question, respondent and interviewer characteristics, and experimental allocations, the impact of changed wording is only significantly associated with question timing; changed wording has a significant negative association with question timing. The other data quality indicators (i.e., Don't Know and distribution of means in the IP

experiments) showed no significant effect from major question-wording deviations. Taken together, although major deviations are significantly associated with shorter question timings, major deviations are not significantly associated with item nonresponse (i.e., Don't Know) or differences in distributions. These results are potentially a positive outcome for researchers using interviewer-administered surveys; major changes to question-wording may not be affecting data quality as researchers think. However, while major deviations may not have the effect we think, our findings suggest other factors affect data quality.

For the data quality measure of 'Don't Know' responses, the findings suggest it is question characteristics that have the greatest impact on data quality. Respondents are about seven times as likely to provide a 'Don't Know answer for sensitive questions than for non-sensitive questions. These results support research that sensitive questions may produce better data quality in self-administered questions (Tourangeau and Smith, 1996). Attitude questions are about twice as likely to have a "Don't Know' response than non-attitude questions. Only a few respondent and interviewer characteristics seem to play a role Don't Know responses, with older respondents are more likely to give "Don't Know" responses, and older interviewers are associated with increases in 'Don't Know'.

In addition to major question-wording changes, a number of question characteristics also affect question timing. Question that appear later in the questionnaire, those which are more difficult, have more words, use of a showcard, attitudinal questions have longer question timing durations, while gate follow-up questions, questions that confirm past information or are double-barrelled have shorter timing durations. As stated previously, we are not saying which is the right time

(shorter or longer) to ensure data quality, but instead to note the importance of the timing differences. This paper focuses on question-wording changes, but these findings of question characteristics and timing should be explored further. Additionally, only respondent age (older) and marital status (single), and interviewer age are significantly associated with longer question times.

Taking into consideration the results of all the data quality indicators, question-wording changes affect question timing. However, for question characteristics, sensitive questions, and attitudinal questions seem to negatively affect data quality as they have an increased risk of Don't Know answers. Among respondent and interviewer characteristics, age seems to play an important factor in data quality. Particularly for the respondent side of the equation, the impact of age on data quality is broadly consistent with differences in cognitive ability (Schwarz and Knauper 1999).

This paper does have limitations. Although we used commonly used data quality measures, the measures do not give a definitive measure of data quality as a validation study would, but instead, show differences (question timing and response distributions) in the data quality measures. For the data quality indicator of 'Don't Know' responses, it may be a slightly better indicator, as most studies treat 'Don't Know' responses as missing data. However, there is some argument that 'Don't Know' responses should be a valid response to some questions. There are also potential issues with the power of our analyses, given the clustering of responses, which may particularly impact the experimental data, which is a subset of a subset.

Conclusion

This thesis examined interviewer question-reading behavior in face-to-face interviewers from several perspectives. Chapter 1 studied the prevalence of interviewer question-reading deviations for face-to-face, fielded surveys, what types of deviations interviewers make, and tested methods for detecting question-reading deviations. Chapter 2 examined how question characteristics may be driving the question-reading deviations. Chapter 3 investigated how question-reading deviations may affect data quality.

To the best of my knowledge, this research is the first study to examine question-reading deviations for *fielded*, *face-to-face* surveys. This distinction is essential as earlier research on these topics uses telephone or lab data where the interviewers can be easily observed and may alter their behavior, compared to face-to-face field interviewers who are largely unobserved.

The research approaches question-reading deviations from three perspectives, 1) interviewer monitoring, 2) questionnaire design, and 3) data quality. Chapter 1 expands the literature by testing previously untested methods (i.e., WPS methods) used by survey organizations to identify potential question-reading deviations. This chapter's research further extends the literature by exploring and testing other methods (i.e., standard deviations and model-based methods) to detect deviations and proposes. Chapter 2 builds on the existing literature by expanding the list of question characteristics used in previous studies, and again is the first study to use behavior coded data and survey data from a fielded face-to-face survey. Likewise, Chapter 3 is the first known study to investigate how question-reading deviations may affect data quality. Further extending the literature, this chapter leverages several IP Wave 3 experiments on question

formation to evaluate whether or not the measurement error found for different question formations can be partially attributed to interviewer question-reading deviations.

The main findings for this research are summarized below:

- Interviewers engaged in major question-reading deviations 13% of the time when administering the survey questions (Chapter 1).
- The question-reading deviations are vastly from interviewers omitting question text (Chapter 1).
- Of the different methods tested to detect question-reading deviations, creating QATTs with the 4WPS method performs the best in terms of accuracy and utility (Chapter 1).
- The research suggests that question characteristics are driving interviewer question-reading deviations. Questions that contain definitions or examples and questions where the response options are read as part of the question have the highest odds of being read with major deviations. (Chapter 2).
- Changed wording has a significant negative association with question timing. Shorter question timings are widely believed to have a negative effect on data quality (Chapter 3).
- The other data quality measures (i.e., Don't Know and distribution of means in the IP experiments) showed no significant effect from major question-wording deviations (Chapter 3).
- While the findings suggest major deviations may not have the negative effect that they are believed to have, caution should be used. This topic is under-researched and requires

> further investigations before we can definitively state that major deviations do not have a negative effect on data quality (Chapter 3).

One of the research aims of this thesis is to provide survey practitioners with recommendations, based on systematic and empirical evidence, on how to best monitor interviewers' behavior during face-to-face interviewers. Using paradata, specifically creating QATTs with a threshold of 4WPS, would allow a targeted, automated approach that should save time and money by reducing the need to listen to all interviews by concentrating quality control efforts on interviews with high rates of questions flagged as having major deviations. This research should also provide insight to questionnaire designers on what types of questions are more likely to induce question-reading deviations and consider including on-screen interview prompts for questions with a higher risk of deviations. Additionally, trainers may want to highlight the questions more prone to deviations during interviewer training and spend more time on the importance of reading all questions verbatim.

This research does fill a gap in the survey research literature. However, since this research is the first study to investigate question-reading deviations for fielded face-to-face interviews, more research is needed. Not only should the studies be replicated, but there are also many directions for future research. For detecting question-reading deviations, future research should consider using a more precise measure of the timing duration. The timing durations used for this study only had times rounded to the nearest second available for analysis. It could be that a more precise time of milliseconds would improve the various methods' accuracy and utility. Another area for future research is testing QATT methods in different languages. As for what is driving question-reading deviations, future research should investigate other respondent and interviewer

characteristics, such as personality traits (e.g., Big 5 Personality traits) that may drive the behavior. For example, it could be that respondents who are lower in agreeableness are more likely to show frustration or respondent burden, and in turn, the interviewer engages more deviations. Similarly, an interviewer lower in conscientiousness may have an increased risk of engaging in question-reading deviations. Also, as stated earlier, to gain a consensus on major deviations and data quality, more research is needed. This study may have potential issues with the analysis's power as a subset of a subset was used for analysis.

Finally, it should be noted that a strength of this thesis is the dataset created by this author to investigate these topics. The combined use of paradata (question timing durations), interviewer behavior coded data, question, respondent and interviewer characteristics data, and survey data made for a rich and rare dataset. This dataset should provide an opportunity to extend the literature on this topic for many years to come.

Bibliography

Ackermann-Piek, D., and Massing, N. (2014). Interviewer behavior and interviewer characteristics in PIAAC Germany. Methods, data, analyses: a journal for quantitative methods and survey methodology (mda), 8(2), 199-222.

Al Baghal, T., & Lynn, P. (2015). Using motivational statements in web-instrument design to reduce item-missing rates in a mixed-mode context. Public Opinion Quarterly, 79(2), 568-579.

Axinn, W. G. (1991). The influence of interviewer sex on responses to sensitive questions in Nepal. Social Science Research, 20(3), 303-318.

Bassili, J. N. and Fletcher, J. F. (1991). Response-Time Measurement in Survey Research a Method for CATI and a New Look at Nonattitudes. Public Opinion Quarterly, 55(3): 331-346.

Belli, R. F. (1998). The Structure of Autobiographical Memory and the Event History Calendar: Potential Improvements in the Quality of Retrospective Reports in Surveys Memory 6, 383-406.

Belli, R. F., Bilgen, I., and Al Baghal, T. (2013). Memory, communication, and data quality in calendar interviews. Public opinion quarterly, 77(S1), 194-219.

Belli, R.F. and Al Baghal, T. (2016). Parallel Associations and the Structure of Autobiographical Knowledge. *Journal of Applied Research on Memory and Cognition*, 5:150-157

Belli, R. F., Lee, E. H., Stafford, F. P., and Chou, C. H. (2004). Calendar and question-list survey methods: Association between interviewer behaviours and data quality. Journal of Official Statistics, 20(2), 185.

Belli, R.F., Miller, L.D., Al Baghal, T , and Soh, L-K (2016). Calendar Interviews: Predicting Respondent Retrieval Strategies. *Journal of Official Statistics*, 32:579-600.

Belli, R. F., and Lepkowski, J. M. (1996, April). Behavior of survey actors and the accuracy of response. In Health Survey Research Methods: Conference Proceedings (pp. 69-74).

Blair, E. (1980). Using practice interviews to predict interviewer behaviors. The Public Opinion Quarterly, 44(2), 257-260.

Bradburn, N. M., Sudman, S., Blair, E., Locander, W., Miles, C., Singer, E., and Stocking, C. (1979). Improving interview method and questionnaire design: Response effects to threatening questions in survey research. University Microfilms.

Cannell, C. F., Lawson, S.A., and Hausser, D.L. (1975). A Technique for Evaluating Interviewer Performance.

Cannell, C.F., Fowler, F.J., and Marquis, K.H. (1968). The Influence of Interviewer and

Respondent Psychological and Behavioral Variables on the Reporting of Household Interviews. Vital and Health Statistics, Series 2, No. 26.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. Sociological methodology, 12, 389-437.

Cannell, C. F., and Robison, S. (1971). Analysis of individual questions. Working papers on survey research in poverty areas, 236-91.

Conrad, F.G., M.P. Couper, R. Tourangeau, and C. Zheng. (2017) Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods* 11:45-61

Couper, M. P. (2000). Usability Evaluation of Computer-Assisted Survey Instruments. Social Science Computer Review, 18(4):384-396.

Couper, M. P., and Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society*), 176: 271-286.

Davern, M., Call, K. T., Ziegenfuss, J., Davidson, G., Beebe, T. J., & Blewett, L. (2008). Validating health insurance coverage survey estimates: a comparison of self-reported coverage and administrative data records. Public Opinion Quarterly, 72(2), 241-259.

Dijkstra, W. (2002). Transcribing, coding, and analyzing verbal interactions in survey interviews. Standardization and Tacit Knowledge, 401-425.

Dykema, J., Lepkowski, J. M., and Blixt, S. (1997). The effect of interviewer and respondent behaviour on data quality: Analysis of interaction coding in a validation study. Survey measurement and process quality, 287-310.

Eckman, S., Kreuter, F., Kirchner, A., Jäckle, A., Tourangeau, R., and Presser, S. (2014). Assessing the mechanisms of misreporting to filter questions in surveys. Public Opinion Quarterly, 78(3), 721-733.

Foulke, E. (1968). Listening comprehension as a function of word rate. Journal of Communication, 18(3), 198-206.

Fowler Jr, F. J., and Cannell, C. F. (1996). Using behavioral coding to identify cognitive problems with survey questions.

Fowler Jr, F. J., and Mangione, T. W. (1990). Standardized survey interviewing: Minimizing interviewer-related error (Vol. 18). Sage.

Gelman, Andrew, and Jennifer Hill. Data analysis using regression and multilevel/hierarchical models. Cambridge university press, 2006.

Gilbert, E. (2015). A Comparison of Branched Versus Unbranched Rating Scales for the Measurement of Attitudes in Surveys. *Public Opinion Quarterly*, 79:443–470.

Goldman, L. E., Chu, P. W., Osmond, D., & Bindman, A. (2011). The accuracy of present-on-admission reporting in administrative data. Health services research, 46(6pt1), 1946-1962.

Green, M. C., Krosnick, J. A., and Holbrook, A. L. (2001). The survey response process in telephone and face-to-face surveys: Differences in respondent satisficing and social desirability response bias. Manuscript, Ohio State University.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2011). Survey methodology (Vol. 561). John Wiley & Sons.

Haan, M., Ongena, Y., and Huiskes, M. (2013). Interviewers' questions: rewording not always a bad thing. Interviewers' deviations in surveys: Impact, reasons, detection and prevention, 173-193.

Holbrook, A., Cho, Y. I., and Johnson, T. (2006). The impact of question and respondent characteristics on comprehension and mapping difficulties. Public Opinion Quarterly, 70(4), 565-595.

Houtkoop-Steenstra, H. (2000). Interaction and the standardized survey interview: The living questionnaire. Cambridge University Press.

Jäckle, A., Gaia, A., Al Baghal, T., Burton J., and Lynn, P. (2017). Understanding Society – The U.K. Household Longitudinal Study, Innovation Panel, Waves 1-9, User Manual. Colchester: University of Essex.

Jans, M. E. (2010). Verbal paradata and survey error: Respondent speech, voice, and question-answering behavior can predict income item nonresponse (Doctoral dissertation, University of Massachusetts Boston).

Jans, M., Sirkis, R., and Morgan, D. (2013). Managing Data Quality Indicators with Paradata Based Statistical Quality Control Tools: The Keys to Survey Performance. Improving Surveys with Paradata, 191-229.

Jones, K., & Subramanian, S. V. (2017). Developing multilevel models for analysing contextuality, heterogeneity and change using MLwiN 3 Volume.

Killpack, C., and Gatenby, R. (2010). Understanding Society Innovation Panel Wave 3: Technical Report. National Centre for Social Research.

Kirgis, N., & Lepkowski, J. M. (2013). Design and Management Strategies for Paradata-Driven Responsive Design: Illustrations from the 2006–2010 National Survey of Family Growth. Improving surveys with paradata: analytic uses of process information, 121-144.

Kirgis, N., Mneimneh, Z., Sun, Y., Lin, Y., & Ndiaye, S. K. (2015). Using paradata to monitor interviewer behavior and reduce survey error. Lecture presented at. In 2015 International Total Survey Error Conference.

Kreuter, F. (2013). Improving surveys with paradata: Introduction. *Improving Surveys with Paradata*, 1-9.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied cognitive psychology, 5(3), 213-236.

Krosnick, J. A., and M. K. Berent. (1993). Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format. *Journal of Political Science* 37:941–64.

Krosnick, J. A., Malhotra, N., and Mittal, U. (2014). Public misunderstanding of political facts: How question wording affected estimates of partisan differences in birtherism. Public opinion quarterly, 78(1), 147-165.

Lenzner, T., L. Kaczmirek, and A. Lenzner (2010), Cognitive Burden of Survey Questions an Response Times: A Psycholinguistic Experiment. *Applied Cognitive Psychology*, 24:1003–1020.

Lepkowski, J. M., Siu, V., and Fisher, J. (2000). Event history analysis of interviewer and respondent survey behavior. Metodoloski Zvezki, 15, 3-20.

Mangione, T. W., Fowler, F. J., and Louis, T. A. (1992). Question characteristics and interviewer effects. Journal of Official Statistics, 8(3), 293.

Marquis, K. H., and Cannell, C. F. (1969). A Study of Interviewer-Respondent Interaction in the Urban Employment Survey. Final Report.

Mathiowetz, N., and Cannell, C. (1980). "Coding Interviewer Behavior as a Method of Evaluating Performance." In Proceedings of the Section on Survey Research Methods, pp. 525-28. Washington, DC: American Statistical Association.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica, 22(3), 276-282.

Mneimneh, Z. N., Pennell, B., Lin, Y., and Kelley, J. (2014). Using paradata to monitor interviewers' behavior: A case study from a national survey in the Kingdom of Saudi Arabia. Comparative Survey Design and Implementation (CSDI) conference

Moore, R. J., and Maynard, D. W. (2002). Achieving understanding in the standardized survey interview: repair sequences. Standardization and tacit knowledge: Interaction and practice in the survey interview, 281-312.

Munzert, S., and Selb, P. (2015). Measuring Political Knowledge in Web-Based Surveys: An Experimental Validation of Visual Versus Verbal Instruments. Social Science Computer Review, 0894439315616325.

Murphy, J., Baxter, R., Eyerman, J., Cunningham, D., and Kennet, J. (2004, May). A system for detecting interviewer falsification. In American Association for Public Opinion Research 59th Annual Conference (pp. 4968-4975).

Oksenberg, L., Cannell, C., and Kalton, G. (1991). New strategies for pretesting survey questions. Journal of official statistics, 7(3), 349.

Olson, K., and Peytchev, A. (2007). Effect of interviewer experience on interview pace and interviewer attitudes. Public Opinion Quarterly, 71(2), 273-286.

Olson, K., and Smyth, J. D. (2015). The effect of CATI questions, respondents, and interviewers on response time. Journal of Survey Statistics and Methodology, 3(3), 361-396.

Omoigui, N., He, L., Gupta, A., Grudin, J., and Sanocki, E. (1999, May). Time-compression: systems concerns, usage, and benefits. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (pp. 136-143). ACM.

Ongena, Y. P. (2005). Interviewer and respondent interaction in survey interviews (Doctoral dissertation, In eigen beheer).

Ongena, Y. P., and Dijkstra, W. (2006a). Methods of behavior coding of survey interviews. Journal of Official Statistics, 22(3), 419.

Ongena, Y. P., and Dijkstra, W. (2006b). Question-answer sequences in survey-interviews. Quality and Quantity, 40, 983-1011. doi: 10.1007/s11135-005-5076-4
Rugg, D. (1941). Experiments in wording questions: II. Public Opinion Quarterly, 5(1), 91.

Peneff, J. (1988). The observers observed: French survey researchers at work. Social Problems, 35(5), 520-535.

Presser, S., and Zhao, S. (1992). Attributes of questions and interviewers as correlates of interviewing performance. The Public Opinion Quarterly, 56(2), 236-240.

Rasinski, K. A. (1989). The effect of question wording on public support for government spending. Public Opinion Quarterly, 53(3), 388-394.

Schaeffer, N.C., and S. Presser. (2003). The Science of Asking Questions. *Annual Review of Sociology* 29:65–88.

Schneider, S. (2018). Extracting response style bias from measures of positive and negative affect in aging research. The Journals of Gerontology: Series B, 73(1), 64-74.

Schober M. F., Conrad F. G. (1997), “Does Conversational Interviewing Reduce Survey

Measurement Error?" Public Opinion Quarterly, 61, 576–602.

Schober, M. F., and Conrad, F. G. (2002). A collaborative view of standardized survey interviews. In D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer and J. van der Zouven (Eds.), Standardization and tacit knowledge: interaction and practice in the survey interview (pp. 67-94). New York, NY: John Wiley and Sons.

Schuman, H., and Presser, S. (1996). Questions and answers in attitude surveys: Experiments on question form, wording, and context. Sage.

Schwarz, N. 1996. *Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation*. Mahwah, New Jersey

Schwarz, N., and Knauper, B. (1999). Cognition, aging, and self-reports. In D. Park and N. Schwarz (Eds.), Cognitive aging—A primer (pp. 233–252). Philadelphia, PA: Psychology Press.

Short, M. E., Goetzel, R. Z., Pei, X., Tabrizi, M. J., Ozminkowski, R. J., Gibson, T. B., ... & Wilson, M. G. (2009). How accurate are self-reports? An analysis of self-reported healthcare utilization and absence when compared to administrative data. Journal of occupational and environmental medicine/American College of Occupational and Environmental Medicine, 51(7), 786.

Steele, F. (2008) Module 5: Introduction to Multilevel Modelling. LEMMA VLE, University of Bristol, Centre for Multilevel Modelling. Accessed at /cmm/lemma.

Sun, Y., and Meng, X. (2014). Using response time for each question in quality control on China Mental Health Survey (CMHS). Comparative Survey Design and Implementation (CSDI) conference

Tang, P. C., Ralston, M., Arrigotti, M. F., Qureshi, L., & Graham, J. (2007). Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures. Journal of the American Medical Informatics Association, 14(1), 10-15.

Thissen, M. R., and Myers, S. K. (2016). Systems and processes for detecting interviewer falsification and assuring data collection quality. Statistical Journal of the IAOS, 32(3), 339-347.

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). The psychology of survey response. Cambridge University Press.

Tourangeau, R., and Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. Public opinion quarterly, 60(2), 275-304.

Uhrig, S. N., and Sala, E. (2011). When change matters: An analysis of survey interaction in dependent interviewing on the British Household Panel Study. Sociological Methods and Research, 40(2), 333-366.

University of Essex, Institute for Social and Economic Research. (2019). Understanding Society: Innovation Panel, Waves 1-11, 2008-2018. [data collection]. 9th Edition. UK Data Service. SN: 6849, http://doi.org/10.5255/UKDA-SN-6849-12.

Van der Zouwen, J., and Dijkstra, W. (2002). Testing questionnaires using interaction coding. Standardization and tacit knowledge: Interaction and practice in the Survey interview, 427-48.

Velez, P., and Ashworth, S. D. (2007, June). The impact of item readability on the endorsement of the midpoint response in surveys. In Survey Research Methods (Vol. 1, No. 2, pp. 69-74).

Viterna, J., and Maynard, D. W. (2002). How uniform is standardization? Variation within and across survey research centers regarding protocols for interviewing. Standardization and tacit knowledge: Interaction and practice in the survey interview.

Wagner, J. (2013). Using Paradata-Driven Models to Improve Contact Rates in Telephone and Face-to-Face Surveys. Improving Surveys with Paradata, 145-1

Wenz, A. (2021). Do distractions during web survey completion affect data quality? Findings from a laboratory experiment. Social Science Computer Review, 39(1), 148-161.

West, B. T., and Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, *5*(2), 175-211

West, B. T., Conrad, F. G., Kreuter, F., and Mittereder, F. (2018). Can conversational interviewing improve survey response quality without increasing interviewer effects?. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(1), 181-203.

Winker, P. (2016). Assuring the quality of survey data: Incentives, detection and documentation of deviant behaviour. Statistical Journal of the IAOS, 32(3), 295-303.

Yan, T. and Tourangeau, R. (2008) Fast times and easy questions: the effects of age, experience and question complexity on web survey response times. Appl. Cogn. Psychol., 22, 51–68.

Zhang C. and Conrad F.G. (2014). Investigation of speeding in web surveys: tendency to speed and its association with straightlining. *Survey Research Methods*, 8:127–135.